Impact Ionisation rate calculations in wide banc gap semiconductors

Harrison, Daniel

Impact Ionisation Rate Calculations in Wide Band Gap Semiconductors

Daniel Harrison

A thesis submitted for the
degree of Doctor of Philosophy
at the University of Durham,
Department of Physics

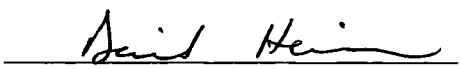
September 1998

Abstract

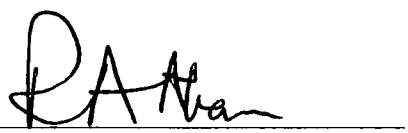
Calculations of band-to-band impact ionisation rates performed in the semi-classical Fermi's Golden Rule approximation are presented here for the semiconductors GaAs, $In_{0.53}Ga_{0.47}As$ and $Si_{0.5}Ge_{0.5}$ at 300K. The crystal band structure is calculated using the empirical pseudopotential method. To increase the speed with which band structure data at arbitrary **k**-vectors can be obtained, an interpolation scheme has been developed. Energies are quadratically interpolated on adapted meshes designed to ensure accuracy is uniform throughout the Brillouin zone, and pseudowavefunctions are quadratically interpolated on a regular mesh. Matrix elements are calculated from the pseudowavefunctions, and include the terms commonly neglected in calculations for narrow band gap materials and an isotropic approximation to the full wavevector and frequency dependent dielectric function. The numerical integration of the rate over all distinct energy and wavevector conserving transitions is performed using two different algorithms. Results from each are compared and found to be in good agreement, indicating that the algorithms are reliable. The rates for electrons and holes in each material are calculated as functions of the **k**-vector of the impacting carriers, and found to be highly anisotropic. Average rates for impacting carriers at a given energy are calculated and fitted to Keldysh-type expressions with higher than quadratic dependence of the rate on energy above threshold being obtained in all cases. The average rates calculated here are compared to results obtained by other workers, with reasonable agreement being obtained for GaAs, and poorer agreement obtained for InGaAs and SiGe. Possible reasons for the disagreement are investigated. The impact ionisation thresholds are examined and **k**-space and energy distributions of generated carriers are determined. The role of threshold anisotropy, variation in the matrix elements and the shape of the bands in determining characteristics of the rate, particularly the softness of the rate's threshold behaviour are investigated.

Declaration

The work presented in this thesis has been carried out by the candidate (except where otherwise acknowledged) and has not been previously submitted for any degree.



Ph.D. Candidate



Ph.D. Supervisor

Acknowledgements

Special thanks go to Dick Abram and Stuart Brand for their supervision during the course of my study, and I am grateful to Alan Beattie for all his assistance. I would like to acknowledge the EPSRC for funding this work, and thank my parents and grandfather for their support, financial and otherwise, especially during my fourth year. Finally, I would also like to thank Mark Walmsley, Dave Hoare, Gavin Crow, Des Ryan, Chris Caulfield, Rick Coles, Dave Dugdale, Steve Pugh and Stewart Clark, without whom this Ph.D. would have probably taken less time, cost less money and been far less enjoyable.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Band-to-band impact ionisation is the process in which an electron in the conduction band collides with an electron in the valence band, exciting it across the band gap and thus creating an electron-hole pair. Alternatively, the process can be initiated by a hole in the valence band colliding with a hole in the conduction band, again resulting in the creation of an electron-hole pair. The process of impact ionisation is associated only with high energy carriers, since the initiating carrier must supply kinetic energy at least equal to the band gap. Significant numbers of such high energy carriers are obtained when created optically by photons with energy well above the band gap, in which case the process of impact ionisation can increase the quantum yield above 1 [1], or more usually when carriers are moving under the influence of a high field. In this latter case, the generated electron-hole pairs are also accelerated by the field and can themselves initiate impact ionisation. If the field is high enough, the resulting auto-catalytic charge multiplication leads to avalanche breakdown [2].

The process of impact ionisation is often detrimental to device performance. The onset of avalanche breakdown above some breakdown voltage (which is characteristic of the material and device structure) limits the maximum output of power devices. Impact ionisation can also be a problem in high speed devices, which are made small to reduce transit times. If the dimensions of the devices are scaled down without a

corresponding reduction in the applied voltages, the internal fields can become high enough to cause significant impact ionisation. This in turn results in undesirable effects such as gate and substrate currents [2–5], and may be responsible for the kink effect in field effect transistors [6–9] and oxide breakdown [10–12]. However, in certain applications impact ionisation can be used to advantage. Detectors such as avalanche photodiodes (APDs) [13,14] and microwave sources such as IMPATT diodes [15] rely on the charge multiplication caused by impact ionisation.

Important quantities in determining the role of impact ionisation in a device are the α and β coefficients, which are defined as the mean number of impact ionisation events initiated by an electron (α) or hole (β) per unit length of drift in the field direction [13,14]. These will generally not be equal, will depend on the material and the field, and can be influenced by the design of the device. Experimentally they can be determined from measurements of gate or substrate currents in FETs [4,5,16] or from multiplication factors in APDs [13,14]. Calculation of the α and β coefficients requires consideration of two aspects of carrier transport: the process by which carriers are accelerated up to sufficiently high energies to initiate impact ionisation, and the rate at which ionisation occurs once carriers have attained sufficient energy. The first of these — the acceleration of carriers to high energy — is commonly studied using Monte Carlo simulation [17–29]. The high energy nature of the impact ionisation process invalidates simple band structure approximations, and the simulations must be carried out using realistic band structure. The resulting numerical complexity requires intensive computational effort. The second aspect — the rate at which the impact ionisation process itself scatters carriers which have attained sufficient energy — is the subject of this thesis. As with the Monte Carlo simulation, the details of the realistic band structure must be considered when calculating the rate, requiring considerable computational resources.

The performance of the simulated device will be affected by several aspects of the impact ionisation scattering rate, which are studied here. Carriers must have at

least some minimum energy to initiate the process. Because the energy distribution of hot carriers is often a rapidly falling function near the threshold energy, the number of carriers in the device able to initiate ionisation will depend sensitively on this threshold energy. For carriers above the threshold, the magnitude of the scattering rate will determine the overall amount of charge multiplication occurring. The performance of avalanche photodiodes depends on the ratio of electron and hole coefficients α/β. Noise in the device can be reduced if α and β differ greatly [13,14], as they do in silicon [30,31], but is increased when $\alpha/\beta \simeq 1$ as is the case in germanium and many III–V materials [30,32]. Thus the relative magnitudes of electron and hole thresholds and rates is of particular interest for such applications. The ratio α/β can be increased (or decreased) and hence the noise performance of APDs improved for III–V materials through the use of heterostructures [14,33–36]. The technique relies on carriers gaining energy at band edge discontinuities rather than through acceleration by the field. The variation of the rate with respect to energy above the threshold is of particular interest. If the rate rises rapidly once threshold is achieved, so that carriers which attain the threshold energy are quickly ionised, the threshold is said to be 'hard', and conversely a slow rise in the rate, allowing carriers to reach energies significantly higher than the threshold energy before ionising, corresponds to a 'soft' threshold. The degree of softness of the threshold can influence various aspects of carrier transport, including the effectiveness of heterostructures in controlling the α and β coefficients [21,37]. Anisotropy of the rate in $\mathbf{k}$-space is of interest as the field-dependence of the α and β coefficients may vary for fields applied in different directions with respect to the crystallographic axes. The coefficients are found to be isotropic in Si [25,31,38] and InP [39,40] for example, while in GaAs there is some disagreement between experiments [41–43], which find them to be anisotropic, and theory [17], which predicts isotropic behaviour. The anisotropy of the rate has implications for the choice of growth direction of devices such as APDs in which the ratio of α/β may vary with direction, or IMPATT diodes in which the avalanche build-up time may vary [44]. Knowledge of the distributions of generated carriers is of

use for performing Monte Carlo device simulations, and of general interest in understanding the process of impact ionisation.

1.1 Previous Work on Impact Ionisation

Early calculations of the impact ionisation α and β coefficients were performed by Wolff [45] and Shockley [46]. Wolff assumed the charge carriers were in thermal equilibrium at a temperature dependent on the applied field, and that impact ionisation was initiated by carriers in the high energy tail of the distribution. Shockley took the opposite approach of assuming that impact ionisation was caused by non-equilibrium 'lucky' electrons which were fortunate enough to avoid collisions and be accelerated to high energy. Wolff's approach is applicable at high fields while Shockley's is applicable at low fields. Baraff [47] assumed that the distribution of electrons was a combination of Wolff's thermalised and Shockley's lucky electrons and obtained results applicable at intermediate fields. More recently, Ridley [48] has developed a simple model, later refined by Burt [49,50], and referred to as the 'lucky-drift' model. It assumes that a carrier's momentum is rapidly randomised by collisions, but that it can escape significant energy relaxation for longer periods. In this way carriers can achieve sufficiently high energy through lucky-drift to initiate impact ionisation. All these theories are concerned mainly with the mechanism by which carriers gain sufficient energy from the field to cause impact ionisation, which is assumed to occur rapidly once the threshold energy is reached. The lucky-drift theory has been refined to take account of a finite scattering rate above threshold [51,52], but does not consider the form of the rate in detail.

Keldysh [53] calculated the actual scattering rate as a function of the energy of the initiating carrier. He applied Fermi's Golden Rule to determine the transition rate due to impact ionisation processes and integrated over all final states. By assuming a direct gap, spherical parabolic band structure and constant transition matrix elements,

he obtained the expression for the rate of ionisation by a carrier at energy E as

$$R(E) = R_0(E - E_0)^P \tag{1.1}$$

where $P = 2$ and R_0 and E_0 are constants dependent on the details of the band structure which, in the application of the model to real materials, are usually fitted to experimental data. Due to its simplicity, the Keldysh formula has been widely used in device simulations, e.g. [17,24,54–56]. However, it is derived using approximations that do not apply to realistic band structure at high energy and so its validity is highly questionable. In addition, it is found that widely varying values of the parameters R_0 and E_0 give similar results when used in Monte Carlo simulations [57], which has resulted in very different scattering rates being used throughout the literature [56], hence giving little physical insight into the process. To determine the role of impact ionisation in devices with more accuracy, the full band structure must be taken into account.

Kane [58] calculated numerically the rate in silicon using realistic band structure, taking into account the proper surfaces of energy and momentum conserving transitions in **k**-space, and the transition matrix elements, thus obtaining the rate as a function of the wavevector (rather than just the energy, as in Eq. (1.1)) of the ionising particle. By neglecting momentum conservation he also derived an energy-dependent expression for the rate, which proved to be a good approximation to that obtained from the full calculation.

Several other authors have obtained the rate using methods very similar to Kane's (e.g. [20,21,59,60]), or a variation in which the Brillouin zone is discretised into small tetrahedra and the rate obtained analytically in each (e.g. [22,25,26]). Beattie has developed an alternative method to calculate the rate involving an explicit surface integration [61], which has been applied to a number of materials [62–64]. All these calculations have several aspects in common:

- The calculations, being based on non-analytic band structure, are very computer-intensive.

- Due to restrictions imposed by energy and crystal momentum conservation, the rate is found to be a function of the specific **k**-vector of the initiating particle [10,22,25,26,65,66] rather than a function of just its energy as in Eq. (1.1). Carriers with the same energy but different wavevectors will in general have widely varying rates.

- For carriers located throughout the Brillouin zone, the individual rates plotted as a function of carrier energy form a scatter graph (as implied in the previous point). If an expression of the form of Eq. (1.1) is fitted through these points, the value of P must usually be set to a value greater than two to obtain the best fit [26,28,67]. This higher P-value is an indication of a softer threshold [22,66], in agreement with the soft threshold indicated by experimental data [51,52,68].

Most recently, calculations [69–72] have been performed which go beyond the semi-classical Fermi's Golden Rule theory, including the effects of collision broadening and intracollisional effect. Both these result in the relaxation of the condition of energy conservation imposed by the semi-classical approach leading to an increase in the anisotropy of the rate and a softening of the thresholds.

1.2 Work Presented in this Thesis

This thesis is concerned with the actual process of impact ionisation and not the transport of carriers up to energies at which it is possible. The materials studied here are GaAs, $In_{0.53}Ga_{0.47}As$ and $Si_{0.5}Ge_{0.5}$, at 300K. GaAs and the alloys $Al_xGa_{1-x}As$ are important materials in the fabrication of high-speed electrical devices and optical devices [73]. The $In_xGa_{1-x}As_yP_{1-y}$ alloys are another important system, of which $In_{0.53}Ga_{0.47}As$ is the lowest band gap composition which is lattice matched to InP. These alloys provide materials for device fabrication with lower band gaps than the $Al_xGa_{1-x}As$ alloys, in particular corresponding to the 1.3 and 1.55μm low dispersion and absorption windows in optical fibres [74], and therefore have applications in optical

communications [75]. Alloys of Si_xGe_{1-x} also have band gaps in the optical fibre low dispersion and absorption windows, and allow the well established Si-based materials technology to participate in the fabrication of devices previously the preserve of the III–V semiconductors [76–78]. The unstrained alloy is considered in this thesis, though in real devices the material may be strained with a corresponding modification in its properties [79,80].

For each material, the calculations carried out here include the following. The impact ionisation thresholds and rates are calculated as a function of wavevector for impacting electrons and holes. The rates are calculated in the semi-classical Fermi's Golden Rule approximation, with the crystal band structure being obtained by the empirical pseudopotential method [81]. An interpolation scheme has been developed to increase the speed with which the band structure data can be retrieved. Matrix elements are obtained from the pseudopotential wavefunctions, including terms commonly neglected in narrow band gap materials [82] and a frequency- and wavevector-dependent dielectric function [83]. The surfaces of allowed transitions are obtained using two different numerical methods: that of Beattie [61] and a method developed here which is a variation of Kane's method [58], optimised to be more efficient near the threshold. Each method has certain strengths and weaknesses, and also the comparison of the two different methods provides a reliable check of the numerical accuracy of the algorithms for the integration over allowed transitions. As well as thresholds and rates, the distributions of the generated carriers are obtained and examined, and the band structure and other factors affecting the rates are investigated.

The rest of this thesis is divided up in the following way. Chapters 2 and 3 deal with the band structure. In Chapter 2 the relevant pseudopotential theory is briefly reviewed, along with the calculation of the dielectric function, which is performed using the method of Walter and Cohen [83]. Chapter 3 covers the implementation of the interpolation scheme used to obtain efficiently the energies, wavefunctions and dielectric function for the rate calculation. Chapters 4 and 5 discuss the impact ionisation process

itself. The basic theory is dealt with in Chapter 4 and details of the implementation of the two integration methods are discussed in Chapter 5. The results are presented in Chapters 6 and 7. General results including thresholds, rates and generated carrier distributions are surveyed in Chapter 6, and the results obtained there are compared to the results of similar calculations performed by other authors. In Chapter 7, a more detailed analysis of the results is performed, with a view to understanding the underlying factors affecting the rates in each material. Finally, in Chapter 8 conclusions are drawn and suggestions for further research made.

Chapter 2

Band Structure Theory

In order to calculate the impact ionisation rate of a carrier in a crystal, a knowledge of the band structure is required — that is to say, we must be able to obtain the energies and wavefunctions of the single electron states throughout the first Brillouin zone. Various methods exist for calculating these quantities, employing greater or lesser approximations, and requiring varying amounts of computational effort.

Whichever method is used, we will normally require the electronic structure information as a function of position in **k**-space, i.e.

$$\begin{aligned} E &= E_n(\mathbf{k}) \\ \psi &= \psi_n(\mathbf{k}) \end{aligned} \tag{2.1}$$

where n is a band index. Frequently we will also require the relation between energy and wavevector in the alternative form $\mathbf{k} = \mathbf{k}_n(E)$, such as when considering the positions of the energy-conserving final states in a given transition. In this case, the positions of $\mathbf{k}$ corresponding to the required energy will have to be searched for, requiring many band structure calculations throughout **k**-space.

2.1 Choice of Calculation Method

Various methods are available for calculating crystal band structure (see, for example, [84], [85]). They can be classified as *ab initio* or *empirical.* Ab initio methods calculate the band structure of the crystal from first principles, requiring few or no adjustable parameters [86]. In contrast, empirical methods rely on several parameters which are adjusted to give results that fit data obtained experimentally. Because of this explicit fitting procedure, the band structure information obtained from empirical methods is generally more reliable (provided the data used is reliable). A further disadvantage of ab initio calculations for this work is the fact that they are usually designed to give ground state properties only — reliable conduction band results cannot be guaranteed.

Since reliable experimental data is available for the semiconductors of interest here, an empirical method is chosen. Furthermore, the method used must be able to provide energy and wavefunction data throughout the Brillouin zone and at high energies, since the electronic states involved in the impact ionisation process will be similarly distributed. This unfortunately rules out the use of effective mass models [87] and the less computationally intensive approaches such as the few band $\mathbf{k}\cdot\mathbf{p}$ method [88]. Instead the method used is the *empirical pseudopotential method* [81,89], which can provide data of acceptable accuracy at the energies required. Fig. 2.1 compares the simple parabolic band approximation for the Γ and satellite valleys with a pseudopotential calculation for the conduction band structure of GaAs. The parabolic approximation is applicable only up to carrier energies of less than an electron volt. Since impact ionisation involves states of much greater energy, the parabolic band approximation is of no use in such calculations.

Unfortunately, the use of the pseudopotential method is considerably more computationally intensive than analytic methods. In fact, in applications such as impact ionisation rate calculations in which the functions $E(\mathbf{k})$ or $\psi(\mathbf{k})$ must be evaluated many times, the pseudopotential method is too CPU intensive to be used directly with

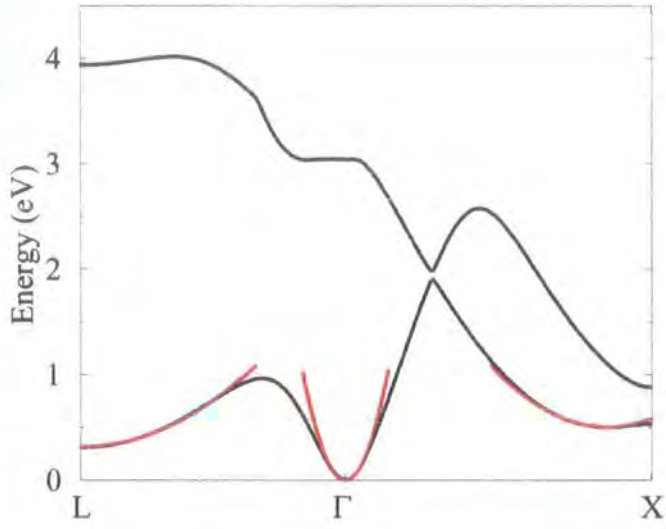


Figure 2.1: The first and second conduction bands of GaAs, obtained by the pseudopotential method (black line) and the parabolic band approximation (red line).

the facilities normally available, and instead is used to set up an interpolation scheme. The implementation of this scheme is the subject of Chapter 3.

2.2 The Pseudopotential Method

The pseudopotential method [81,89–91] can provide energy and wavefunction data as a function of position in **k**-space throughout the Brillouin zone and at all energies of interest here. The method seeks to solve the Schrödinger equation numerically for single electron states in the crystal, but with the real crystal potential replaced by a *pseudopotential* which is weaker, but nevertheless gives the same energy band structure. To see why the pseudopotential is required in practical numerical calculations, consider attempting to solve the Schrödinger equation using the real potential.

2.2.1 Direct Solution of the Hamiltonian

The allowed energies E and eigenfunctions $\psi(\mathbf{r})$ of single electron states in the crystal are obtained by solving the Schrödinger equation

$$\left(\frac{\hat{p}^2}{2m} + V(\mathbf{r})\right)\psi(\mathbf{r}) = E\psi(\mathbf{r}) \tag{2.2}$$

where $V(\mathbf{r})$ is a function representing the average potential felt by each electron due to the lattice of ions and the other electrons. If we know the form of $V(\mathbf{r})$ we can in principle solve Eq. (2.2) for all possible E and $\psi(\mathbf{r})$. The potential is expanded as a Fourier series in terms of reciprocal lattice vectors, $\mathbf{G}$:

$$V(\mathbf{r}) = \sum_m V(\mathbf{G}_m)e^{i\mathbf{G}_m\cdot\mathbf{r}}. \tag{2.3}$$

For a given point in $\mathbf{k}$-space, eigenfunctions of the crystal Hamiltonian can be written in the Bloch form, i.e. as $\psi_\mathbf{k}(\mathbf{r}) = u_\mathbf{k}(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}$ where $u_\mathbf{k}(\mathbf{r})$ has the periodicity of the crystal and can be expanded (like the potential) as a Fourier series. N plane waves are used in the expansion, giving

$$\psi_\mathbf{k}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}\sum_{n=1}^{N} a_n(\mathbf{k})e^{i\mathbf{G}_n\cdot\mathbf{r}}. \tag{2.4}$$

where a sufficiently large value of N must be chosen to ensure that the corresponding energy eigenvalue converges with respect to it.

The potential in Eq. (2.3) and the wavefunction in the form of Eq. (2.4) are put into Eq. (2.2), which is then reduced to a matrix eigenvalue problem.

$$\begin{pmatrix} T_1 & V_{12} & \cdots & V_{1N} \\ V_{21} & T_2 & \cdots & V_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ V_{N1} & V_{N2} & \cdots & T_N \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix} = E\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix} \tag{2.5}$$

The elements of the matrix are

$$T_i = \frac{\hbar^2}{2m}(\mathbf{k} + \mathbf{G}_i)^2 \tag{2.6}$$

$$V_{ij} = \langle \mathbf{K}_i | V(\mathbf{r}) | \mathbf{K}_j \rangle \tag{2.7}$$

where $|\mathbf{K}_n\rangle = e^{i(\mathbf{k}+\mathbf{G}_n)\cdot\mathbf{r}}$.

Using N plane waves in the expansion of the wavefunction at given $\mathbf{k}$, we obtain N eigenvalues, each corresponding to the energy of a different band at $\mathbf{k}$. For each eigenvalue there is an eigenvector whose components give us the coefficients $a_1 \ldots a_N$ for the wavefunction of the band in question at $\mathbf{k}$. Thus solving Eq. (2.2) at given $\mathbf{k}$ gives energies and wavefunctions for as many bands as we use plane waves in the expansion of the wavefunction[a].

In practice this method cannot be used. The rapid oscillations of the wavefunction in the regions of large negative potential around the ions of the lattice requires the expansion of Eq. (2.4) to contain of the order of 10^6 terms or more to achieve a good (i.e. converged) representation. Solving the eigenvector problem obtained from putting this into the Schrödinger equation thus requires diagonalisation of matrices of $10^6 \times 10^6$ elements — the computational requirements for this are clearly prohibitive.

2.2.2 The Pseudo-Hamiltonian

Pseudopotential techniques seek to find an alternative form of the potential which gives the same eigenvalues (band energies), but leads to eigenvectors (wavefunctions) that require expansion in terms of fewer coefficients. The matrices to be diagonalised are thus smaller and the computational requirements much reduced.

The pseudopotential method assumes that the electronic states can be divided into two types: *core* states and *valence* states. The core states are the closed shells of inner electrons around each ion. The core states of neighbouring ions do not overlap, and

[a]Note that only the lowest solutions can be assumed to be converged — the N^{th} eigenfunction for example will certainly not be.

the wavefunctions of the core states of each ion are the same as in the free atoms. (The energy levels in the core will all be shifted due to the coulomb interaction with neighbouring ions, but the inter-level energy separations are assumed to remain the same). The valence states are the remaining outer states. Here 'valence' denotes any non-core state i.e. the states known as the valence and conduction bands in the usual semiconductor terminology. It is these states (and not the core states) that influence the properties of the crystal of interest here.

The lowest eigenvector obtained from a direct solution of Eq. (2.2) would be the $1s$ core state. The remaining core states — $2s$, $2p$, etc ... — would come next, then finally the valence states. In the pseudopotential method used here, the lowest eigenvector corresponds to the first valence state — the core states are 'by-passed'. This is achieved by expanding the wavefunction in a basis set which is itself orthogonal to the core states — a set of *orthogonalised plane waves.*

Orthogonalised Plane Waves

An orthogonalised plane wave (OPW) is a plane wave with core states added in such a way as to make it orthogonal to those core states [90]:

$$\mathrm{OPW}(\mathbf{k}) = |\mathbf{k}\rangle - \sum_{\alpha} |\alpha\rangle\langle\alpha|\mathbf{k}\rangle \tag{2.8}$$

$$\langle \mathrm{OPW}|\alpha\rangle = 0 \tag{2.9}$$

where $|\mathbf{k}\rangle = e^{i\mathbf{k}\cdot\mathbf{r}}$ is a plane wave and $|\alpha\rangle$ is an atomic core wavefunction. We then express the electron wavefunction as an expansion in terms of M orthogonalised plane waves,

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{j=1}^{M} b_j(\mathbf{k}) \left(1 - \sum_{\alpha} |\alpha\rangle\langle\alpha|\right)|\mathbf{k} + \mathbf{G}_j\rangle \tag{2.10}$$

where the $b_1 \ldots b_M$ are coefficients to be determined. Eq. (2.10) is substituted into Eq. (2.2) and the resulting terms re-arranged to give

$$\left(\frac{\hat{p}^2}{2m} + V(\mathbf{r}) + \sum_{\alpha}(E_{\mathbf{k}} - E_c)|\alpha\rangle\langle\alpha|\right)\varphi_{\mathbf{k}}(\mathbf{r}) = E\varphi_{\mathbf{k}}(\mathbf{r}) \tag{2.11}$$

where $\varphi_{\mathbf{k}}(\mathbf{r}) = \sum_j b_j(\mathbf{k})e^{i(\mathbf{k}+\mathbf{G}_j)\cdot\mathbf{r}}$ is a sum of plane waves. Finally, Eq. (2.11) is written as

$$\left(\frac{\hat{p}^2}{2m} + \widehat{V}_{ps}\right)\varphi_{\mathbf{k}}(\mathbf{r}) = E(\mathbf{k})\varphi_{\mathbf{k}}(\mathbf{r}). \tag{2.12}$$

This pseudo-Hamiltonian is of the same form as the original Schrödinger equation Eq. (2.2), and can be solved in the same way. However the real potential $V(\mathbf{r})$ has been replaced by a non-local operator called the *pseudopotential*,

$$\widehat{V}_{ps} = V(\mathbf{r}) + \sum_{\alpha}(E_{\mathbf{k}} - E_c)|\alpha\rangle\langle\alpha| \tag{2.13}$$

and the real wavefunction $\psi_{\mathbf{k}}(\mathbf{r})$ has been replaced by a *pseudowavefunction* $\varphi_{\mathbf{k}}(\mathbf{r})$, related to the real one by

$$\psi_{\mathbf{k}}(\mathbf{r}) = \left(1 - \sum_{\alpha}|\alpha\rangle\langle\alpha|\right)\varphi_{\mathbf{k}}(\mathbf{r}) \tag{2.14}$$

Eqs. (2.2) and (2.12) have the same energy eigenvalues, except that eigenvalues corresponding to the core states are only obtained from the real Hamiltonian and not from the pseudo-Hamiltonian. Because of the initial expansion of the wavefunction in a basis orthogonal to the core states, the lowest eigenvalue of the pseudo-Hamiltonian corresponds to the first valence state.

2.2.3 The Advantage of the Pseudopotential

The important difference between the real potential $V(\mathbf{r})$ and the pseudopotential $\widehat{V}_{ps}$ is that the pseudopotential is *weak* in the sense that it has no bound states in the region of the ionic cores. This is due to the fact that the second term on the right

hand side of Eq. (2.13) acts to cancel out the effect of the first term (the real attractive potential). This in turn results in the pseudowavefunctions being smooth functions throughout all space, including near the ion cores where the real wavefunctions oscillate rapidly. This smooth characteristic of the pseudowavefunctions allows them to be expanded in terms of only a few plane waves ($\sim$ 100). The matrices that require diagonalisation in order to solve the pseudo-Hamiltonian equation are correspondingly small, and the computational requirements more manageable. Note however that for many applications the computational requirements are not reduced to the point where the pseudopotential calculation can be used directly. As will be discussed in Chapter 3, in this work it is used indirectly through an interpolation scheme.

Figs. 2.2 and 2.3 are schematic diagrams comparing the real potential and wavefunction with their pseudo- counterparts. In Fig. 2.2 it can be seen that the real potential is very deep near the ion itself ($\sim \frac{1}{r}$). In contrast, the pseudopotential remains weak near the ion. Similarly, the real wavefunction oscillates rapidly near the ion while the pseudowavefunction remains smooth. Away from the ion — that is, outside the volume occupied by the inner core orbitals — the potential and wavefunction are the same in both real and pseudo cases.

In many applications the pseudowavefunctions can be used as approximations to the real wavefunctions without converting from one to the other via Eq. (2.14). This is due to the fact that they only differ in the small volume of the core regions — see §2.3.2.

2.3 Solving the Pseudo-Hamiltonian

The pseudo-Hamiltonian, Eq. (2.12), is solved in the same way as described at the beginning of §2.2 for the real Hamiltonian, Eq. (2.2). The pseudopotential $\widehat{V}_{ps}$, and the Bloch part of the pseudowavefunction $\varphi(\mathbf{r})$ are each expanded as a Fourier series in terms of reciprocal lattice vectors $\mathbf{G}$, as was done in Eqs (2.3) and (2.4) for the

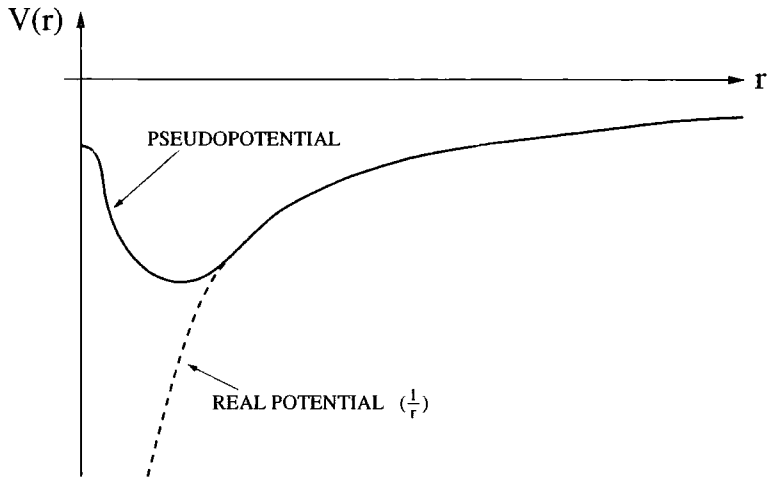


Figure 2.2: A schematic diagram of the pseudopotential (solid line) and real ionic potential (dashed line). The real potential becomes very strong near the ion ($V \sim \frac{1}{r}$) but the pseudopotential remains weak everywhere. Away from the ion, pseudo- and real potentials are the same.

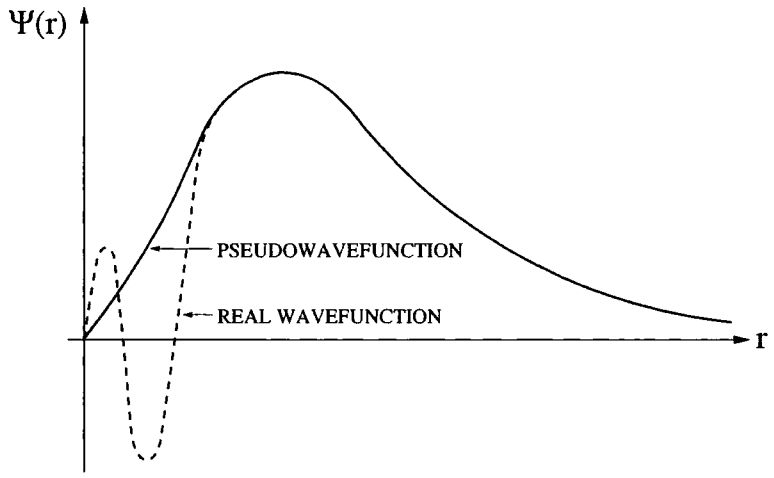


Figure 2.3: A schematic diagram of the pseudowavefunction (solid line) and real wavefunction (dashed line). The real wavefunction oscillates rapidly near the ion (where the real potential is strong) whereas the pseudowavefunction has no such rapid variation. Away from the ion, pseudo- and real wavefunctions are the same.

real potential and wavefunction. These are put into the pseudo-Hamiltonian, leading to an eigenvector problem of the form of Eq. (2.5) whose solution gives the energies and pseudowavefunctions. The problem then is to evaluate the matrix elements, given by Eqs. (2.6) and (2.7) in the case of the real potential, which in the case of the pseudopotential become

$$T_i = \frac{\hbar^2}{2m}(\mathbf{k}+\mathbf{G}_i)^2 + \langle \mathbf{K}_i|\widehat{V}_{ps}|\mathbf{K}_i\rangle \tag{2.15}$$

$$V_{ij} = \langle \mathbf{K}_i|\widehat{V}_{ps}|\mathbf{K}_j\rangle \tag{2.16}$$

Calculating the matrix elements using the expression for $\widehat{V}_{ps}$ in Eq. (2.13) would be a complicated task, and require knowledge of the core states $|\alpha\rangle$. Cohen and Bergstresser [89] approximated $\widehat{V}_{ps}$ with a simple local potential $V_L(\mathbf{r})$ (which is much weaker than the real potential), and obtained a good fit to the experimentally determined band structurefor a number of semiconductors. Chelikowsky and Cohen [81] improved the accuracy of the fitted band structure by including non-local terms in the pseudopotential, as well as terms accounting for the spin-orbit interaction. It is this form of the pseudopotential which is used in this work. The discussion here will concentrate on the parameters needed to specify the potential, and for a full account the reader should refer to [81].

Using the pseudopotential of [81], the matrix elements V_{ij} can be considered to be the sum of three parts:

$$\langle \mathbf{K}_i|\widehat{V}_{ps}|\mathbf{K}_j\rangle = \langle \mathbf{K}_i| \;\; V_L + \widehat{V}_{NL} + \widehat{V}_{SO} \;\; |\mathbf{K}_j\rangle \tag{2.17}$$

in which V_L is a local pseudopotential, $\widehat{V}_{NL}$ is a non-local pseudopotential and $\widehat{V}_{SO}$ accounts for spin-orbit coupling. Each of these terms is described below.

The Local Pseudopotential, V_L

The first term in the pseudopotential is a local potential, i.e. a simple function of position, $\mathbf{r}$. The local potential is expanded as a Fourier series in terms of reciprocal

lattice vectors, $\mathbf{G}$:

$$V_L(\mathbf{r}) = \sum_n V(\mathbf{G}_n) e^{i\mathbf{G}_n \cdot \mathbf{r}} \tag{2.18}$$

For the diamond and zinc-blende structures it is convenient to express the $V(\mathbf{G})$'s in the form

$$V(\mathbf{G}) = V^S(\mathbf{G}) \cos(\mathbf{G} \cdot \boldsymbol{\tau}) + iV^A(\mathbf{G}) \sin(\mathbf{G} \cdot \boldsymbol{\tau}) \tag{2.19}$$

where $\boldsymbol{\tau} = \frac{1}{8}a_0(1,1,1)$, a_0 being the lattice constant, and where V^S and V^A are the *symmetric* and *antisymmetric form factors* for the crystal. They are related to the form factors for the potential due to the cation and anion by

$$\begin{aligned} V^S &= \frac{1}{2}(V^c + V^a) \\ V^A &= \frac{1}{2}(V^c - V^a). \end{aligned} \tag{2.20}$$

If we assume the pseudopotentials due to the cation and anion are spherically symmetric, $V^S(\mathbf{G})$ and $V^A(\mathbf{G})$ become functions only of the magnitude of $\mathbf{G}$ i.e. $V^S(G)$ and $V^A(G)$.

For the crystals of interest here, converged band structure can usually be obtained using form factors up to and including the $G = \sqrt{11}\frac{2\pi}{a_0}$ reciprocal lattice vectors (i.e. the $(3,1,1)$ vectors). This gives us three symmetric and three antisymmetric form factors[b]. In the case of the diamond structure, the antisymmetric form factors must all be zero due to the fact that the same ion is located at each site in the basis.

Thus the local part of the pseudopotential is specified in terms of the parameters $V^S(\sqrt{3})$, $V^S(\sqrt{8})$, $V^S(\sqrt{11})$, $V^A(\sqrt{3})$, $V^A(\sqrt{4})$, $V^A(\sqrt{11})$ and a_0.

The Non-Local Pseudopotential, $\widehat{V}_{NL}$

The second term in the pseudopotential of Eq. (2.17) is a non-local potential. For each of the cation and anion, the non-local potential is written as a sum of spherical

[b]The $V^S(\sqrt{4})$ and $V^A(\sqrt{8})$ form factors do not contribute due to $\cos(\mathbf{G} \cdot \boldsymbol{\tau})$ and $\sin(\mathbf{G} \cdot \boldsymbol{\tau})$ being zero for the $\mathbf{G}$-vectors with these magnitudes respectively.

potential wells surrounding the ion, each of which acts on a different angular momentum component of the wavefunction. Matrix elements for each ion are then of the form

$$\langle \mathbf{K}_i|\widehat{V}_{NL}|\mathbf{K}_j\rangle = \sum_l \langle \mathbf{K}_i|A_l(E)f_l(r)\widehat{\mathcal{P}}_l|\mathbf{k}_j\rangle \tag{2.21}$$

The sum over l includes all angular momentum states present in the core wavefunctions, i.e. for the semiconductors of interest here, the s, p and d states with $l = 0, 1$ or 2. For one of these components the non-local well can be 'absorbed' into the expression for the local potential, and as in [81] this is done for the p-well. Thus the sum is over the components $l = 0$ and $l = 2$. The terms in the sum are as follows.

$A_l(E)$ is the well depth and in general is a weak function of energy. The s-well depth is given by

$$A_0(E) = \alpha_0 + \beta_0 \left\{ \left[E^0(K_i)E^0(K_j)\right]^{\frac{1}{2}} - E^0(K_F) \right\} \tag{2.22}$$

and it is adequate to give the d-well a fixed depth, A_2.

The function $f_l(r)$ defines the shape of the well, which is usually taken to be a square well[c]:

$$f_l(r) = \begin{cases} 1 & r < R_l, \\ 0 & r > R_l \end{cases} \tag{2.23}$$

in which R_l is some suitable value for the radius.

The term $\widehat{\mathcal{P}}_l$ is the projection operator for the l^{th} angular momentum component of the wavefunction, which ensures that the l^{th} well acts only on the appropriate component of the wavefunction.

The non-local potential has therefore introduced five new parameters for each of the cation and anion: α_0, β_0, R_0, A_2 and R_2, i.e. ten new parameters in total.

[c]The square well is used due to its simplicity, although in GaAs for example, a gaussian well of suitable width is used.

Spin-Orbit Coupling, $\widehat{V}_{SO}$

The final term in the pseudopotential in Eq. (2.17) accounts for the effect of the spin-orbit interaction. Matrix elements for the cation and anion are of the form

$$\begin{aligned}\langle \mathbf{K}_i|\widehat{V}^c_{SO}|\mathbf{K}_j\rangle &= -\mu i\langle \nu_i|\widehat{\boldsymbol{\sigma}}|\nu_j\rangle \cdot (\mathbf{K}_i \times \mathbf{K}_j)\\ \langle \mathbf{K}_i|\widehat{V}^a_{SO}|\mathbf{K}_j\rangle &= -\alpha\mu i\langle \nu_i|\widehat{\boldsymbol{\sigma}}|\nu_j\rangle \cdot (\mathbf{K}_i \times \mathbf{K}_j)\end{aligned} \tag{2.24}$$

where ν_n is a spinor and the components of $\widehat{\boldsymbol{\sigma}}$ are the Pauli spin matrices. In the Chelikowsky and Cohen formulation, the spin-orbit matrix elements also include terms of the form

$$B_{nl}(K) \propto \int_0^\infty j_l(Kr)\ R_{nl}(r)\ r^2 \mathrm{d}r \tag{2.25}$$

where $j_l(Kr)$ is a spherical Bessel function and $R_{nl}(r)$ is a radial wavefunction. In this work, the contribution of these terms is found to be negligible in InGaAs and SiGe, but is included in GaAs (for which the relevant data was already available).

Inclusion of the spin-orbit interaction in the pseudo-Hamiltonian doubles the number of terms used to expand the pseudowavefunction. Where previously there were N terms corresponding to N reciprocal lattice vectors, there are now $2N$ terms: N spin-up terms and N spin-down. The size of the matrix involved in the resulting eigenvector problem is correspondingly doubled. (An alternative approach, which is used by Chelikowsky and Cohen but not used in this work, is to include the spin-orbit interaction as a perturbation which is applied after the matrix eigenvector problem has been solved [92,93]. This leads to wavefunction expansions of $2N$ terms as before, but avoids the doubling in size of the Hamiltonian matrix).

Inclusion of the spin-orbit interaction introduces two new adjustable parameters for fitting: μ and α, corresponding to the overall strength of the interaction and the cation-anion weighting respectively.

2.3.1 Fitting Pseudopotentials

The pseudopotential used here is described in terms of 19 parameters. Since the amount of experimental data available for fitting these parameters may well be limited for a given material, it is undesirable to use so many adjustable parameters in fitting the band structure. Therefore, as in the procedure of [81], several are fixed before any fitting is carried out. The non-local potential radii can all be given fixed values. The s-well radii (R_0) are fixed using the Heine-Animalu [94,95] calculations, while the d-well radii are are set to $R_2 = \frac{\sqrt{3}}{8}a_0$, which makes the d-wells touching spheres. The value of α appearing in the spin-orbit term is set at the ratio of the spin-orbit splittings in each of the free atoms, and of course the lattice constant a_0 is known beforehand. Thus the fitting procedure must adjust the values of 13 of the parameters to fit the pseudopotential band structure to experimental data. Table 2.1 summarises the situation.

In this work, adjustment of the 13 fitted parameters was performed by a Monte Carlo fitting method which compares the band structure of the pseudopotential calculation to the experimental data, and randomly adjusts each parameter, until satisfactory agreement is achieved. The procedure and further constraints that can be applied to the fitting parameters are discussed in Chapter 3, §3.1.

2.3.2 Output of the Pseudopotential Calculation

Once the parameters defining the pseudopotential have been set, the pseudopotential calculation itself (that is, the solving of the pseudo-Hamiltonian) can be viewed as a 'black box'. It takes as input one **k**-vector in the 1st Brillouin zone and gives as output energies and wavefunctions for the first $2N$ bands[d] at that **k**-vector, where N is the number of plane waves used to expand the wavefunction.

[d]The factor of 2 is due to the inclusion of spin.

Parameter Symbol	Parameter	Fitted/Fixed
$V_S(\sqrt{3})$ $V_S(\sqrt{8})$ $V_S(\sqrt{11})$	Symmetric form factors	
$V_A(\sqrt{3})$ $V_A(\sqrt{4})$ $V_A(\sqrt{11})$	Antisymmetric form factors	Fitted
α_0^c, β_0^c α_0^a, β_0^a	s-well depths (cation & anion)	
A_2^c A_2^a	d-well depths (cation & anion)	
μ	Spin-orbit coupling	
R_0^c R_0^a	s-well radii (cation & anion)	
R_2^c R_2^a	d-well radii (cation & anion)	Fixed
α	Ratio of S-O splitting in free atoms	
a_0	Lattice constant	

Table 2.1: The parameters required to specify the form of the pseudopotential. Those marked 'Fitted' are adjusted to give band structure fitting experimentally measured data. Those marked 'Fixed' can be set before fitting takes place.

Energies and Wavefunctions

Energies are output in the form of $2N$ scalars

$$\text{Energies} \quad = \quad E_n(\mathbf{k}), \quad n = 1 \ldots 2N \tag{2.26}$$

where $E_1(\mathbf{k})$ is the energy of the lowest valence state at $\mathbf{k}$, and energy increases with increasing band index n — see Table 2.2. Wavefunctions are output as $2N$ vectors

$$\text{Wavefunctions} \quad = \quad \mathbf{c}_n(\mathbf{k}), \quad n = 1 \ldots 2N \tag{2.27}$$

where vector $\mathbf{c}_n$ corresponds to the state with energy E_n, and each vector has $2N$ components,

$$\mathbf{c}_n(\mathbf{k}) = \Big(\quad {}^{\uparrow}c_{n,1}(\mathbf{k}), \ldots, {}^{\uparrow}c_{n,N}(\mathbf{k}), \quad {}^{\downarrow}c_{n,1}(\mathbf{k}), \ldots, {}^{\downarrow}c_{n,N}(\mathbf{k}) \quad \Big) \tag{2.28}$$

each of which are generally complex numbers. The pseudowavefunction $\varphi_n(\mathbf{k})$ is obtained from the vector $\mathbf{c}_n(\mathbf{k})$ using the expression

$$\varphi_n(\mathbf{k}) \equiv u_n(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{r}} = \frac{1}{\sqrt{\Omega}} \left[\sum_{j=1}^{N} \Big({}^{\uparrow}c_{n,j}|\uparrow\rangle + {}^{\downarrow}c_{n,j}|\downarrow\rangle \Big) e^{i\mathbf{G}_j\cdot\mathbf{r}} \right] e^{i\mathbf{k}\cdot\mathbf{r}} \tag{2.29}$$

where $u_n(\mathbf{k})$ is the Bloch periodic part of $\varphi_n(\mathbf{k})$ and Ω is the volume of the crystal. The eigenfunctions $|\uparrow\rangle$ and $|\downarrow\rangle$ are orthonormal spin-up and spin-down states respectively. Thus, a general wavefunction is not a pure spin-up or spin-down eigenstate but is a linear combination of the two.

Fig. 2.4 shows the first 20 energy bands of GaAs obtained using the pseudopotential calculation. Note that each line on the plot is in fact a pair of bands, which along L–Γ and Γ–X are degenerate, but along X–U, K–Γ and at general points in the Brillouin zone are minutely split by the spin-orbit interaction. These pairs will frequently be referred to together, for example, the expression '1$^{\text{st}}$ conduction band' will be used to denote the first pair of bands immediately above the band gap. Table 2.2 lists the band indices (as returned by the pseudopotential calculation) along with the usual names

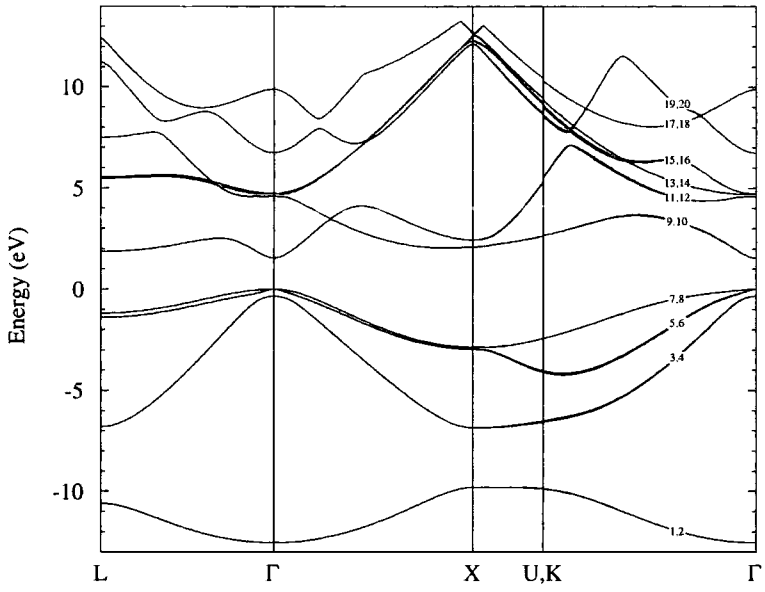


Figure 2.4: The lowest 20 energy bands of GaAs, obtained by the pseudopotential method. Along L–Γ–X the bands are doubly degenerate. In the other directions, each line on the plot is in fact pair of bands, minutely split by the spin-orbit interaction. The numbering on the right hand side indicates the band indices. See also Table 2.2.

given to those bands in semiconductors.

Overlap Integrals, Orthonormality and Degeneracy

The overlap integral between the periodic parts of Bloch functions at general values of wavevector is given by the expression

$$\langle u_m(\mathbf{k}_1)|u_n(\mathbf{k}_2)\rangle = \sum_{j=0}^{N} {}^{\uparrow}c^*_{m,j}(\mathbf{k}_1)\ {}^{\uparrow}c_{n,j}(\mathbf{k}_2) + {}^{\downarrow}c^*_{m,j}(\mathbf{k}_1)\ {}^{\downarrow}c_{n,j}(\mathbf{k}_2). \tag{2.30}$$

The wavefunctions of the bands output for a given wavevector form an orthogonal set which can be normalised, and hence

$$\langle u_m(\mathbf{k})|u_n(\mathbf{k})\rangle = \delta_{m,n} \tag{2.31}$$

where $\delta_{m,n}$ is the Kronecker delta function.

Calculations of crystal properties such as the impact ionisation rate require the evaluation of overlap integrals of the form of Eq. (2.30). The true matrix element should

Band Index	Name
3,4	Spin Split-off
5,6	Light Hole
7,8	Heavy Hole
9,10	1^{st} Conduction
11,12	2^{nd} Conduction
...	...

Table 2.2: The band indices of the pseudopotential calculation (with spin) and their usual names in semiconductors. See also Fig. 2.4.

be calculated using the real wavefunctions, obtained from the pseudowavefunctions via Eq. (2.14). However the additional core-core and plane wave-core terms introduced by using the real instead of pseudo wavefunction are small due to the small volume occupied by the core, and it is usually sufficient to use pseudowavefunctions in the evaluation of matrix elements [96].

At a general point in **k**-space in a zinc-blende semiconductor, all bands will be non-degenerate, but this is not necessarily the case at symmetry points. At the Γ-point for example, all bands are at least doubly degenerate as the spin-orbit interaction does not cause the small splitting between pairs of bands that it does elsewhere. The wavefunctions of degenerate bands are not uniquely specified as at non-degenerate points. Any linear combination of degenerate wavefunctions is equally valid, and the exact combination output by the pseudopotential 'black box' is in general random. However, the degenerate wavefunctions output from the calculation can be combined to ensure that they are orthogonal to each other and normalised, just as in the non-degenerate case.

In the case of diamond structure semiconductors, this spin-degeneracy is present at all points in **k**-space and hence no single-electron state has a uniquely defined wavefunction.

2.4 The Dielectric Function

The expression for the impact ionisation matrix element includes the longitudinal dielectric function ϵ of the crystal [97] (see §4.2 of Chapter 4), which is calculated from the band structure obtained using the pseudopotential method. Generally ϵ is a function of wavevector and frequency of the field being screened, $\epsilon = \epsilon(\mathbf{q}, \omega)$, and is obtained in this work from the expression [83]

$$\epsilon(\mathbf{q},\omega) = 1 + \frac{e^2}{\Omega\epsilon_0 q^2} \sum_{\mathbf{k},c,v} \left| \left\langle u^c_{\mathbf{k}} \left| u^v_{\mathbf{k}+\mathbf{q}} \right\rangle \right|^2 \times \left\{ [E_c(\mathbf{k}) - E_v(\mathbf{k}+\mathbf{q}) - \hbar\omega - i\eta]^{-1} + [E_c(\mathbf{k}) - E_v(\mathbf{k}+\mathbf{q}) + \hbar\omega + i\eta]^{-1} \right\}. \tag{2.32}$$

where Ω is the crystal volume, $E_n(\mathbf{k})$ and $u^n_{\mathbf{k}}$ are the energy and Bloch periodic part of the wavefunction at $\mathbf{k}$ in the n^{th} band, η is a positive infinitesimal value and the sum is over all $\mathbf{k}$-states in the 1^{st} Brillouin zone and all valence bands v and conduction bands c.

This expression has real and imaginary parts which, writing them explicitly as ϵ_r and ϵ_i respectively are

$$\epsilon_r(\mathbf{q},\omega) = 1 + \frac{e^2}{\Omega\epsilon_0 q^2} \sum_{\mathbf{k},c,v} \left| \left\langle u^c_{\mathbf{k}} \left| u^v_{\mathbf{k}+\mathbf{q}} \right\rangle \right|^2 \times \left\{ [E_c(\mathbf{k}) - E_v(\mathbf{k}+\mathbf{q}) - \hbar\omega]^{-1} + [E_c(\mathbf{k}) - E_v(\mathbf{k}+\mathbf{q}) + \hbar\omega]^{-1} \right\}. \tag{2.33}$$

and

$$\epsilon_i(\mathbf{q},\omega) = \frac{\pi e^2}{\Omega\epsilon_0 q^2} \sum_{\mathbf{k},c,v} \left| \left\langle u^c_{\mathbf{k}} \left| u^v_{\mathbf{k}+\mathbf{q}} \right\rangle \right|^2 \delta(E_c(\mathbf{k}) - E_v(\mathbf{k}+\mathbf{q}) - \hbar\omega) \tag{2.34}$$

The real and imaginary parts can be calculated directly from these expressions. In each case the sum over $\mathbf{k}$ can be performed by Monte Carlo sampling of the 1^{st} Brillouin zone, and in the case of evaluation of the imaginary part, the Dirac delta function can approximated by a top-hat function of small energy width and unit area. The numerical evaluation of ϵ_r and ϵ_i is discussed further in Chapter 3, §3.5.

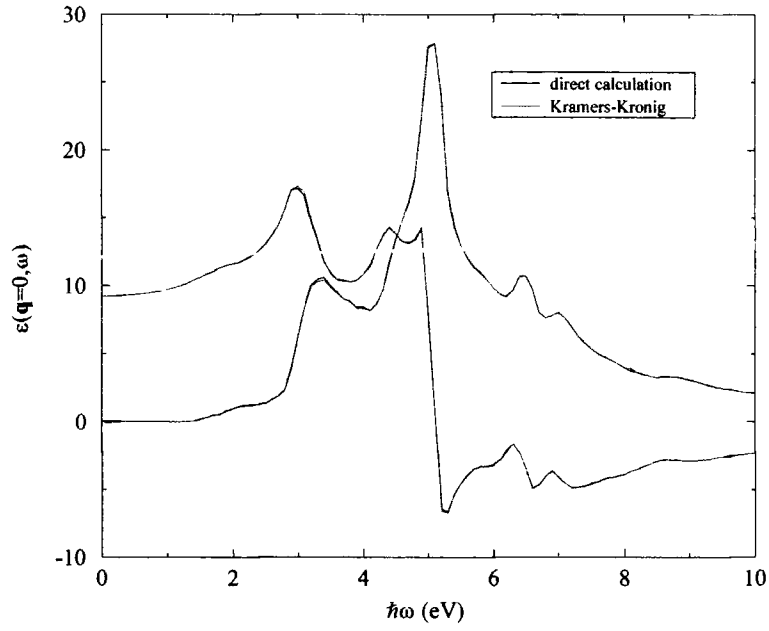


Figure 2.5: The real and imaginary parts of the dielectric function of GaAs, as a function of frequency at fixed wavevector ($\mathbf{q} = \mathbf{0}$). The black lines correspond to ϵ_r and ϵ_i calculated directly using the pseudopotential method. The red lines correspond to ϵ_r obtained through the use of the Kramers-Kronig relations from the pseudopotential calculation of ϵ_i, and vice versa.

The real and imaginary parts of the dielectric function are also related by the Kramers-Kronig expressions [97]

$$\epsilon_r(\mathbf{q},\omega) = 1 + \frac{2}{\pi}\int_0^\infty \frac{\omega'\epsilon_i(\mathbf{q},\omega')}{\omega'^2 - \omega^2}d\omega' \tag{2.35}$$

$$\epsilon_i(\mathbf{q},\omega) = -\frac{2\omega}{\pi}\int_0^\infty \frac{\epsilon_r(\mathbf{q},\omega') - 1}{\omega'^2 - \omega^2}d\omega' \tag{2.36}$$

If both parts of the dielectric function have been calculated, these relations can be used as a test of the numerical accuracy of the calculation. Alternatively, if only the real(imaginary) part has been calculated, the imaginary(real) part can be obtained through the use of Eqs. (2.35) and (2.36). Fig. 2.5 shows the real and imaginary parts of the dielectric function of GaAs, calculated directly using Eqs. (2.33) and (2.34), and via the Kramers-Kronig relations. The black and red lines are almost indistinguishable except for small differences near $\hbar\omega \simeq 3$eV, indicating good numerical accuracy.

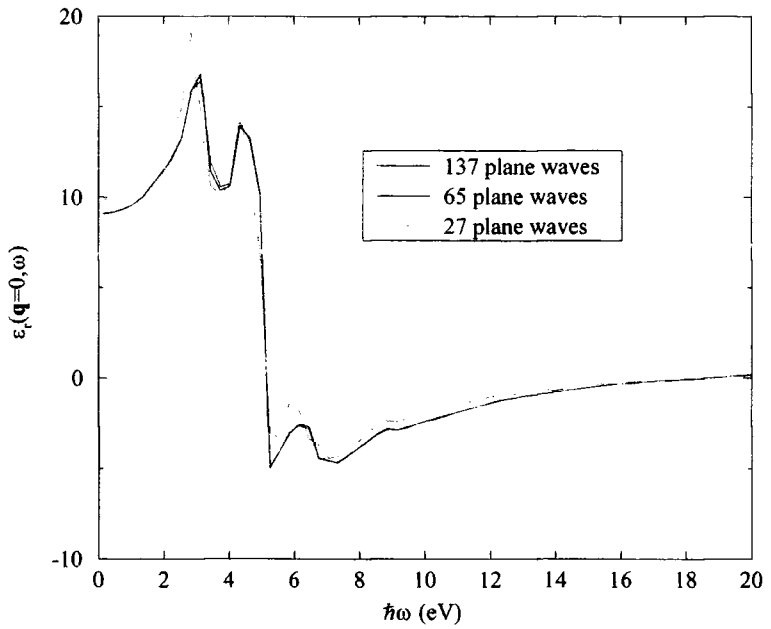


Figure 2.6: The convergence of the dielectric function with respect to the number of plane waves used in the pseudopotential calculation.

Convergence of the Dielectric Function

The basis set of plane waves used to expand the pseudowavefunctions is finite. Use of the pseudopotential rather than the real crystal potential ensures that convergence of the wavefunction is rapid with respect to the number of plane waves used and Fig. 2.6 gives an illustration of this. It shows the result of calculating the real part of $\epsilon(\mathbf{q} = 0, \omega)$ for GaAs using 27, 65 and 137 plane waves in the expansion. From the plot it can be seen that convergence is very good by 65 plane waves (as used in this work).

Chapter 3

Interpolation Schemes

To carry out calculations of crystal properties such as impact ionisation rates in semiconductors, it is necessary to have information on the band structure, in particular the single electron state energies $E(\mathbf{k})$ and wavefunctions $\psi(\mathbf{k})$ at all values of wavevector $\mathbf{k}$ in each of the relevant bands. The empirical pseudopotential method discussed in Chapter 2 is well suited to this task since it can provide accurate band structure information throughout the Brillouin zone for many bands above and below the fundamental band gap. However, the method is CPU intensive and the number of pseudopotential calculations required during execution of a typical impact ionisation rate calculation would take an impractical length of time and cannot be used directly.

To overcome this problem, we make use of the storage of pre-calculated information and *interpolation*. Rather than perform many band structure calculations within the application, data previously obtained for chosen bands and positions in $\mathbf{k}$-space by the pseudopotential method is used. We interpolate $E(\mathbf{k})$ and $\psi(\mathbf{k})$ at arbitrary $\mathbf{k}$ from the stored information. Performing the interpolation is more rapid than the full pseudopotential calculation by several orders of magnitude. While the initial pre-calculation is time consuming, it only needs to be performed once. Each time the application is run, the interpolation scheme reduces execution time to a manageable level.

Unfortunately, the use of an interpolation scheme inevitably incurs a loss of accuracy in the band structure obtained. For an arbitrary **k**, the energy and wavefunction interpolated from the stored values will not exactly match that which would have been obtained with a direct pseudopotential calculation. Of course, the pseudopotential method itself has inherent errors, but it is important to limit as much as possible the further error introduced by the interpolation scheme.

The interpolation error, i.e. the discrepancy between interpolated and calculated band structure data, depends on the density of pre-calculated points in **k**-space. If calculated points are closely spaced in the Brillouin zone[a], arbitrary **k**-vectors will never lie far from a stored point and interpolation errors will be low. Conversely, if the pre-calculated points are sparsely distributed, interpolation errors may far exceed any error inherent in the pseudopotential method itself. However, the demands on the computer's memory must also be considered. Thus, we must design an interpolation scheme with a satisfactory combination of accuracy, memory-efficiency and rapid data retrieval.

3.1 Pre-Calculation of Band Structure — Fitting

The empirical pseudopotential method of calculating band structure relies on several adjustable parameters, which are chosen to give the best possible fit to experimental band structure data for the material in question. These parameters are listed in Table 2.1 of Chapter 2. As explained in the caption below that table, the parameters labelled 'Fixed' are given values independently of the experimental data fitted to. The remaining parameters — those marked 'Fitted' — are adjusted in this work by a Monte Carlo method which compares the band structure obtained from the pseudopotential calculation with the experimental data, and randomly adjusts each parameter, until satisfactory agreement is achieved.

[a] As will be discussed in §3.2, points need only be distributed throughout a small part of the Brillouin zone.

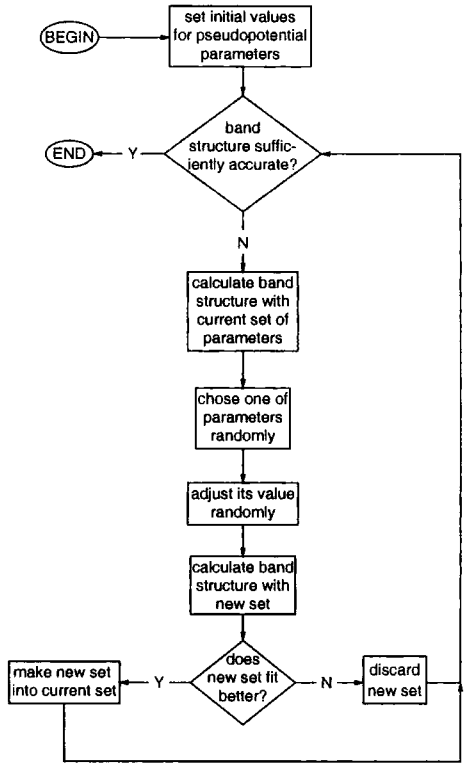


Figure 3.1: The Monte Carlo algorithm used to adjust parameters for the pseudopotential calculation so as to fit the experimentally determined band structure.

The fitting error e is calculated using the expression

$$e = \sum_i w_i (E_i - F_i)^2 \tag{3.1}$$

where E_i and F_i are the experimental and fitted values of the i^{th} energy difference, and w_i is a weighting factor applied to the i^{th} difference. This weighting allows energy differences for which particularly reliable experimental data is available, such as the fundamental band gap, to be fitted more accurately at the expense of less important values such as those relating to certain higher conduction bands. The aim of the fitting procedure is to minimise the value of e in Eq. (3.1) by adjusting the pseudopotential parameters. The algorithm is represented in Fig. 3.1.

Although the adjustment of the pseudopotential parameters is random, certain constraints can be placed on the values they take, as follows:

- The final fitted band structure is required with spin-orbit effects included. However, these effects are relatively small and the fitting procedure is performed without them. Energies are fitted taking into account the splittings that will

occur when the effects are later included, e.g. for fitting purposes, the band gap is taken to be the real (with spin) band gap E_g, plus $\frac{1}{3}$ of the spin split off gap Δ_0.

- The spin parameter μ (see Table 2.1) is determined after the Monte Carlo fitting procedure is finished, and is chosen so as to ensure that the spin split off gap at the top of the valence band matches the experimental value.

- In non-elemental semiconductors, the anti-symmetric form factors are constrained to be a monotonically decreasing function of **G**-vector magnitude. In elemental semiconductors, the anti-symmetric form factors are always zero.

- In the non-elemental semiconductors, the remaining non-local parameters can take different values for the anion and cation. In the elemental semiconductors, in which the atoms of the basis are the same, 'anion' and 'cation' values for the non-local parameters are constrained to be the same.

Fig. 3.2 shows the result of an example fit for $Si_{0.5}Ge_{0.5}$. The energy gaps for which experimental data was used for fitting are indicated. Table 3.1 shows how closely the calculated band structure reproduces the experimental results. The fundamental gap, E_g was given the highest weighting during fitting, resulting in it being the best fit. The vertical gaps at Γ are also well fitted — all to within 50 meV. The gaps at X and L were given the least weighting during the fit due to the unreliable experimental information for these, and as a result are the worst fitted.

When suitable values for the pseudopotential parameters have been obtained using the fitting procedure, they can be used to pre-calculate the interpolation data.

3.2 The Irreducible Wedge

The interpolation scheme is required to provide band structure data at points throughout the first Brillouin zone. However, we can make use of the symmetry of the reciprocal

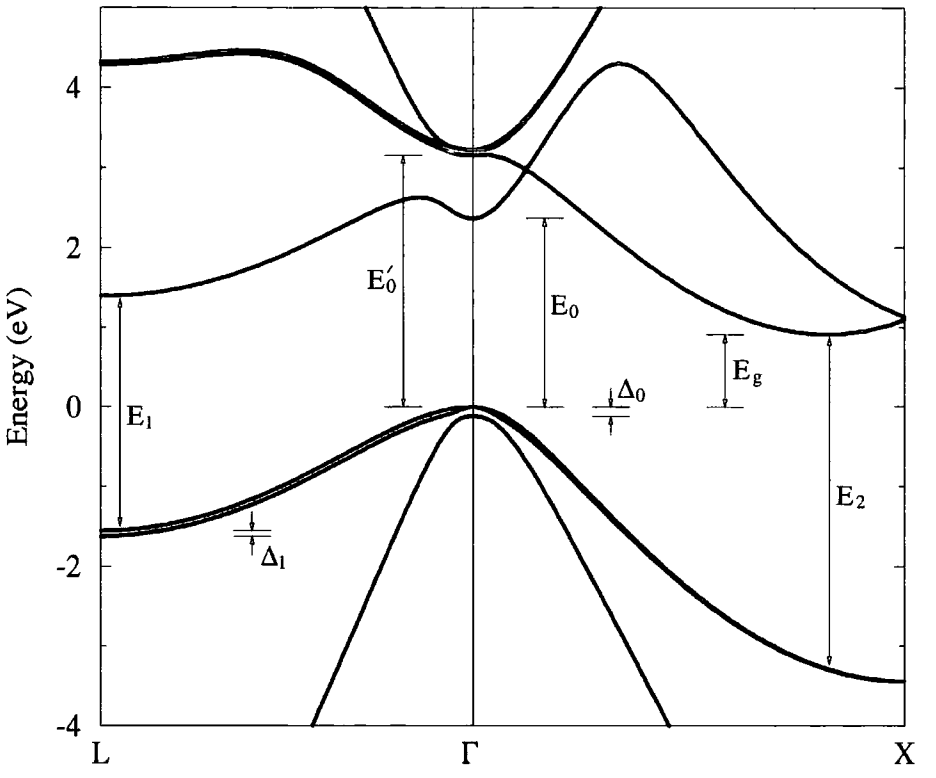


Figure 3.2: The fitted band structure of $Si_{0.5}Ge_{0.5}$ at room temperature. The experimentally measured energy gaps which are used for fitting the pseudopotential parameters are indicated.

Gap	Experiment (eV)	Fit (eV)
E_g (Minimum gap)	0.905	0.908
E_0	2.409	2.360
$E_0 + \Delta_0$	2.524	2.475
E_1	2.724	2.944
$E_1 + \Delta_1$	2.866	3.012
$E'_0 + \Delta_0$	3.242	3.265
E_2	4.358	4.194

Table 3.1: Comparison of energy gaps for $Si_{0.5}Ge_{0.5}$, determined from experiment [98] at room temperature and from the fitted pseudopotential calculation. See Fig. 3.2 for the position of the energy gaps.

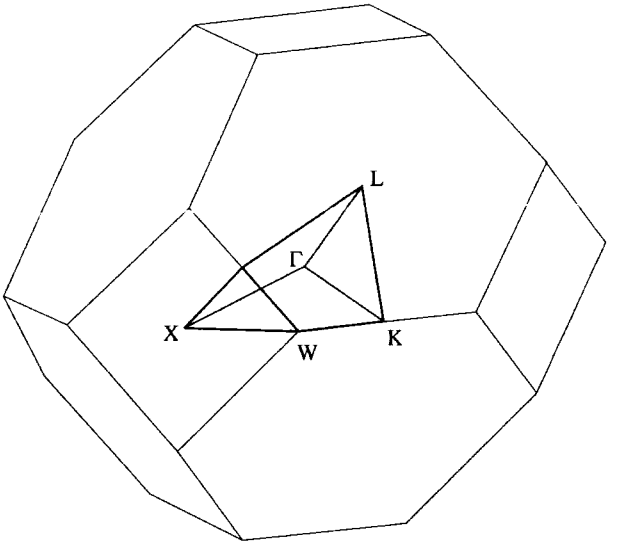


Figure 3.3: The Brillouin zone of the Zinc Blende lattice, and its irreducible wedge. The special points Γ(000), X(100), L($\frac{1}{2}\frac{1}{2}\frac{1}{2}$), W($1\frac{1}{2}0$) and K($\frac{3}{4}\frac{3}{4}0$), all in units of $\frac{2\pi}{a_0}$, are shown.

lattice to reduce the amount of data we actually need to store.

The Brillouin zone can be divided into 48 equal wedge-shaped volumes. One such wedge is shown in Fig. 3.3. It occupies the volume defined (in units of $\frac{2\pi}{a_0}$) by

$$\begin{aligned} 0 \leq k_z \leq k_y \leq k_x \leq 1 \\ k_x + k_y + k_z \leq 1.5 \end{aligned} \tag{3.2}$$

and is known as the *irreducible wedge.* Each point $\mathbf{k}_1$ in the irreducible wedge has a 'corresponding' point $\mathbf{k}_n$ in the n^{th} wedge ($n = 2 \ldots 48$) which can obtained by permuting the coordinates of $\mathbf{k}_1$ (6 permutations) and changing their signs (a further 8 combinations, giving 48 wedges in total).

These permutations and combinations are performed by the application of symmetry operations which are listed in Table 3.2 with the corresponding transformations of coordinates. Thus, if we know the band structure throughout the volume of the irreducible wedge, we can obtain the band structure at any point the Brillouin zone by the application of the appropriate symmetry operations. Both energy and wavefunction data can be obtained in this way.

Before	Operation	After
(k_x, k_y, k_z)	120° rotation about [111]	(k_z, k_x, k_y)
(k_x, k_y, k_z)	240° rotation about [111]	(k_y, k_z, k_x)
(k_x, k_y, k_z)	180° rotation about [100]	$(k_x, -k_y, -k_z)$
(k_x, k_y, k_z)	180° rotation about [010]	$(-k_x, k_y, -k_z)$
(k_x, k_y, k_z)	180° rotation about [001]	$(-k_x, -k_y, k_z)$
(k_x, k_y, k_z)	reflection in $k_x = k_y$ plane	(k_y, k_x, k_z)
(k_x, k_y, k_z)	time inversion	$(-k_x, -k_y, -k_z)$

Table 3.2: The set of symmetry operations required to transform a **k**-point from one irreducible wedge to any other.

To illustrate this point, Fig. 3.4 shows a contour map of the 1st conduction band of GaAs, plotted in the $k_z = 0$ plane. From the plot it can be seen that the $E(\mathbf{k})$ relation in each of the eight irreducible wedges is the same, to within a suitable transformation of coordinates. Thus to obtain energy at arbitrary **k**, it is necessary only to know the energy at the corresponding **k** within the irreducible wedge.

By making use of this symmetry property of the Brillouin zone we are able to reduce the amount of band structure data it is necessary to store by a factor of 48, which is a crucial saving as the accuracy of our interpolation scheme will ultimately be limited by the amount of RAM available to store the data.

3.3 Energy Interpolation

Electron energies in the crystal are a function of **k** and band index b: $E = E_b(\mathbf{k})$. Each band is interpolated separately, so for an electron in a given band we must interpolate energy as a function of three Cartesian **k**-space coordinates: k_x, k_y and k_z.

Quadratic interpolation is used. This was found to be considerably better than the simpler linear interpolation due to the nearly parabolic nature of the band structure at the band extrema.

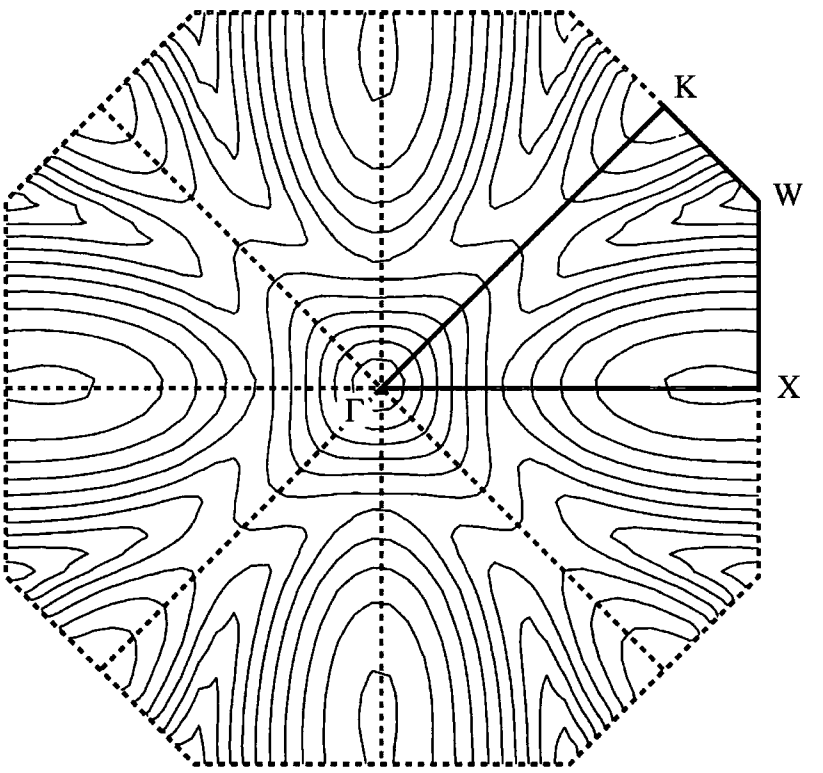


Figure 3.4: The 1st conduction band of GaAs, plotted in the $k_z = 0$ plane. Note that the energy in the irreducible wedge (marked by a solid outline) can be used to obtain the energy in any other wedge.

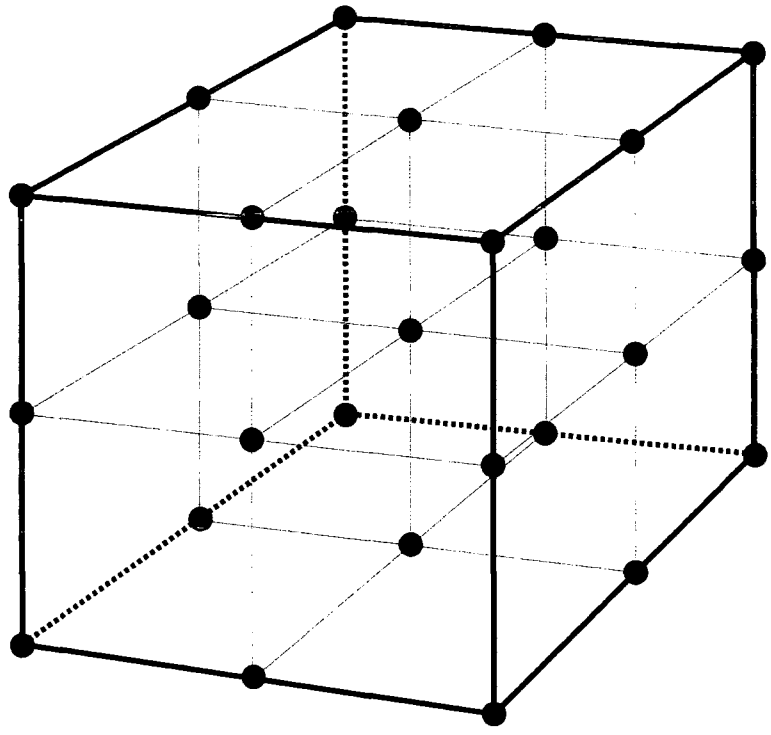


Figure 3.5: An interpolation element. The energy at a **k**-point within its volume is interpolated using the energies stored at the 27 nodes, marked as solid circles.

3.3.1 Implementing the Interpolation Scheme

The irreducible wedge is divided into cubic regions of space, which will be known as interpolating *elements*. Each element has associated with it 27 **k**-points, or *nodes*, at each of which the energy of the particular band in question is stored. Fig. 3.5 shows a cubic interpolating element and its 27 nodes.

Other interpolation schemes (e.g. [99]) have used tetrahedral interpolating elements. Cubic elements are chosen for this work primarily to improve the interpolation scheme's ease of implementation and retrieval of band structure information.

A polynomial of the form

$$E(k_x, k_y, k_z) = e_1 + e_2 k_x + e_3 k_y + e_4 k_z + e_5 k_x^2 + \ldots$$
$$\cdots + e_{25} k_x^2 k_y k_z^2 + e_{26} k_x^2 k_y^2 k_z + e_{27} k_x^2 k_y^2 k_z^2 \qquad (3.3)$$

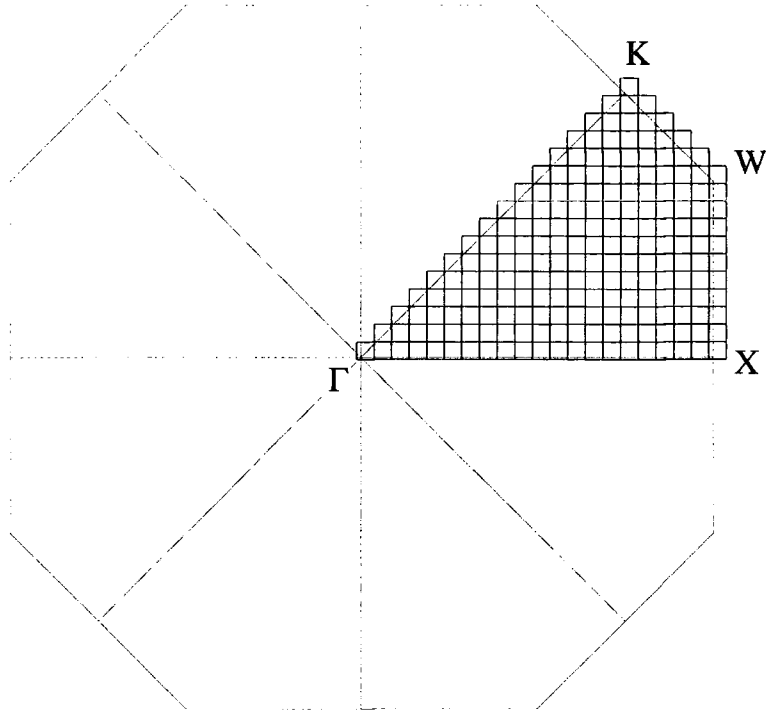


Figure 3.6: A regular interpolation grid's intersection with the $\mathbf{k}_z = 0$ plane. Note how the grid extends beyond the volume of the irreducible wedge — to allow interpolation right up to the edge of the zone.

is used to approximate the energy throughout the volume of the interpolating element. The polynomial contains 27 constants, $e_1 \ldots e_{27}$, which are set so that the energy given by the pseudopotential calculation is obtained at the nodes.

The interpolation elements fill the volume of the irreducible wedge, as shown in Fig. 3.6. Any arbitrary point in the irreducible wedge will therefore be contained within one of these elements. In order to interpolate band structure right up to the boundary of the wedge, the elements extend just outside. Thus some of the pre-calculated band structure lies on nodes not actually within the irreducible wedge. Note that, although each element has 27 nodes, elements share nodes thus reducing the number of **k**-points stored.

To obtain energy at an arbitrary point in **k**-space, the element containing that point must first be identified, and then used to interpolate the energy at the exact **k**-point in question. The algorithm is represented in Fig. 3.7.

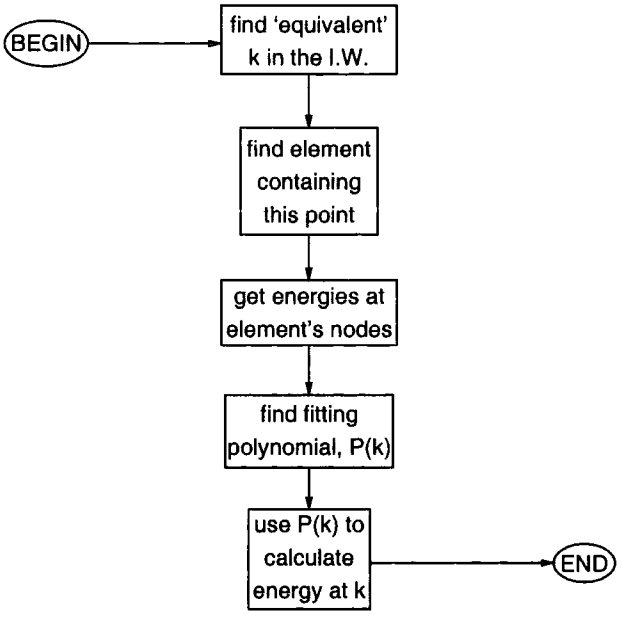


Figure 3.7: The energy interpolation algorithm.

3.3.2 Adapted Grids

As stated earlier, the accuracy of the interpolation is determined mainly by the density of pre-calculated **k**-points (nodes) in the region to be interpolated. We would ideally like as many points throughout the irreducible wedge as possible, thus improving the accuracy, but are limited by the amount of available computer memory. A useful compromise is to concentrate the interpolating elements in regions of the band structure which are difficult to interpolate, leaving the more easily interpolated regions with fewer elements. Such an 'adapted' grid for the lowest conduction band of GaAs is shown in Fig. 3.8.

A comparison of Figs. 3.4 and 3.8 shows that the elements are clustered around the region between the Γ- and X-valleys, which is a rapidly varying part of the band structure and thus difficult to interpolate. In contrast the smoothly varying and nearly parabolic region around the X-valley minimum contains fewer elements, due to the ease of interpolation here.

Generally a different adapted grid is required for each band, tailored to the specific

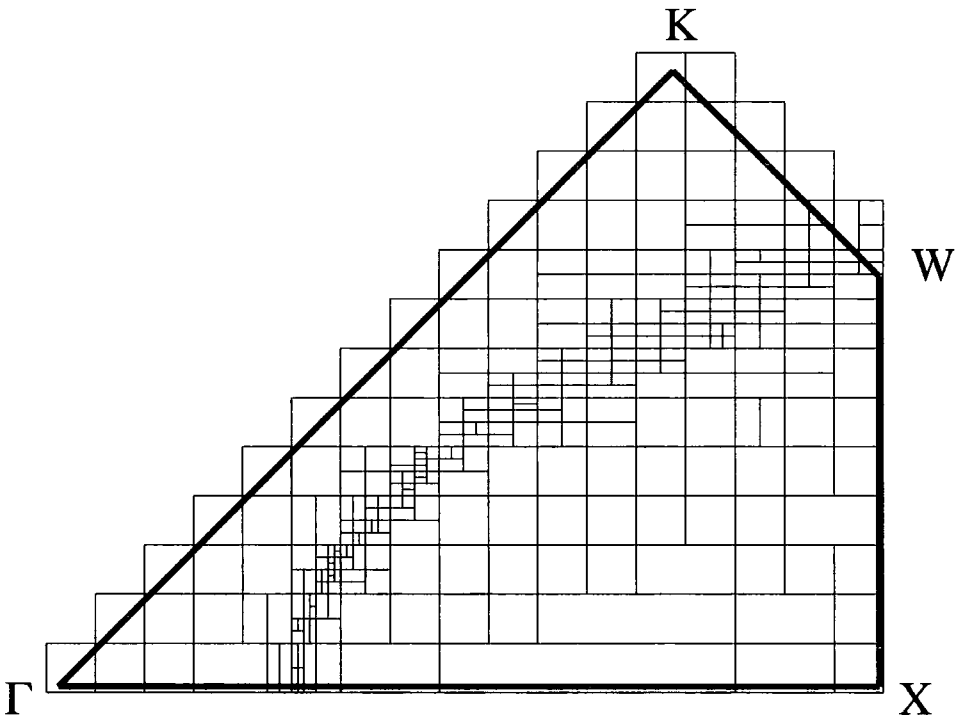


Figure 3.8: An adapted interpolation grid for the 1st conduction band of GaAs. The interpolating points are most densely packed into the region between the Γ- and X-valleys where the energy varies rapidly — compare with Fig. 3.10.

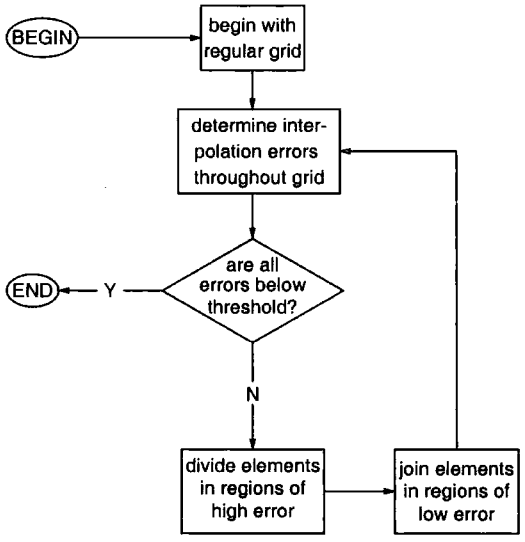


Figure 3.9: The adaptation algorithm. The aim is to distribute the grid points so as to ensure interpolation errors are uniform throughout the irreducible wedge, and below some threshold value.

shape of that band. The grid is produced by starting with a simple regular grid such as the one in Fig. 3.6. In regions where the interpolation is inaccurate, the elements can be divided in one or more of the k_x, k_y or k_z directions to form two, four or eight new elements. Alternatively, pairs of elements lying in easily interpolated regions can be joined together to form a single element. This process is repeated several times for each element until the inaccuracy of the interpolation has been reduced to some predetermined threshold throughout the irreducible wedge. In this work the threshold was taken to be 4 meV. There is no advantage to improving the accuracy beyond this as the band structure produced by the pseudopotential method is itself only of this order of accuracy. The algorithm is represented in Fig. 3.9

3.3.3 Quality of the Interpolation

The aim of the interpolation is to provide accurate band structure as rapidly as possible, working within the limitations of the available computer memory. In this section, the scheme's performance in these three respects is examined.

Memory Use

The interpolation scheme is implemented in Fortran 77, with all data stored to single precision — that is to say one real number requires four bytes of storage space. Thus, four bytes are required to store a single energy value, and 12 bytes are required to store a **k**-vector. Additional memory is used to store the information specifying how these **k**-vectors are combined in groups of 27 to form elements. Table 3.3 shows the memory required by various grids constructed for GaAs.

Band	Number of **k**-points	RAM required (MBytes)
1–8	6 127	0.311
9,10	9 969	0.274
11,12	25 695	0.718
13,14	45 595	1.270
Total	87 386	2.573

Table 3.3: Memory requirements of stored energies for GaAs, based on the assumption that a real number can be stored in four bytes of RAM.

The valence bands (bands 1–8) use the same grid, which needs to be stored only once. For higher bands, a new grid needs to be used for each pair[b] of bands, adapted to their specific features. The number of points (nodes) needed to obtain the required accuracy increases rapidly for the higher bands, where the structure becomes more complicated.

The total memory required to store all the energy data for GaAs is under three megabytes — well within the capability of a modern workstation.

Speed

The time required to perform a large number of interpolations or calculations was measured using a Hewlett-Packard 735 workstation. It was found that the direct pseu-

[b]Each band, e.g. the first conduction band, is in fact a pair of nearly degenerate bands which at a general **k**-point are split by the spin-orbit interaction. These bands are very similar in shape, and one grid can be adapted for both.

dopotential eigenvalue calculation could be performed at a rate of 3.1 calculations per second (i.e. energy band structure information could be obtained at 3.1 positions in **k**-space per second). In comparison, 7000 interpolations could be performed each second — an increase in speed by a factor of 2250.

Accuracy

The accuracy of the interpolation scheme is tested by choosing a large number of **k**-vectors randomly throughout the irreducible wedge and at each comparing the energies evaluated by interpolation and by the full calculation. Figs. 3.10 and 3.11 compare the interpolation errors incurred using a regular grid of 9177 **k**-points, and an adapted grid of 9969 **k**-points respectively.

Both are plotted in the $k_z = 0$ plane of the Brillouin zone. In each case the vertical axis measures the energy as a function of **k**, while the colour denotes the degree of interpolation error. It can be seen that using the regular grid, error increases around the tightly curved (and non-parabolic) region of band structure between the Γ- and X-valleys. With the adapted grid, errors are reduced to a roughly uniform level throughout the wedge. Table 3.4 summarises the interpolation error levels band-by-band.

Band	RMS interpolation error (meV)
1–8	0.96
9,10	0.85
11,12	0.89
13,14	0.93

Table 3.4: Interpolation errors for GaAs — that is, the RMS difference between the interpolated and calculated values for the energy evaluated at a large number of randomly chosen points throughout the zone.

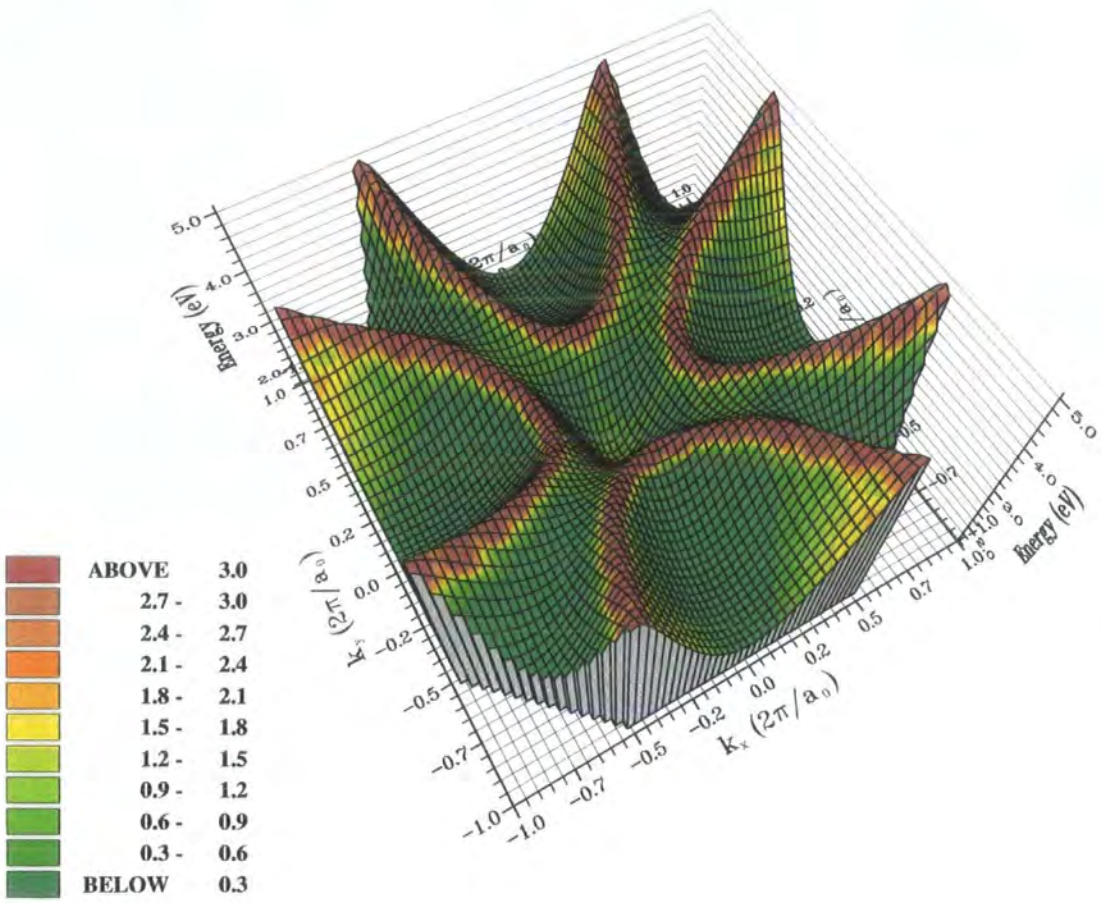


Figure 3.10: Interpolation errors on a regular grid. The base of the plot is the $\mathbf{k}_z = 0$ plane of the Brillouin zone and the height denotes the 1st conduction band energy. The plot is coloured according to the interpolation error at that point: the key is in units of meV. Note that the worst error occurs in the region between the Γ- and X-valleys — compare with the distribution of points in Fig. 3.8.

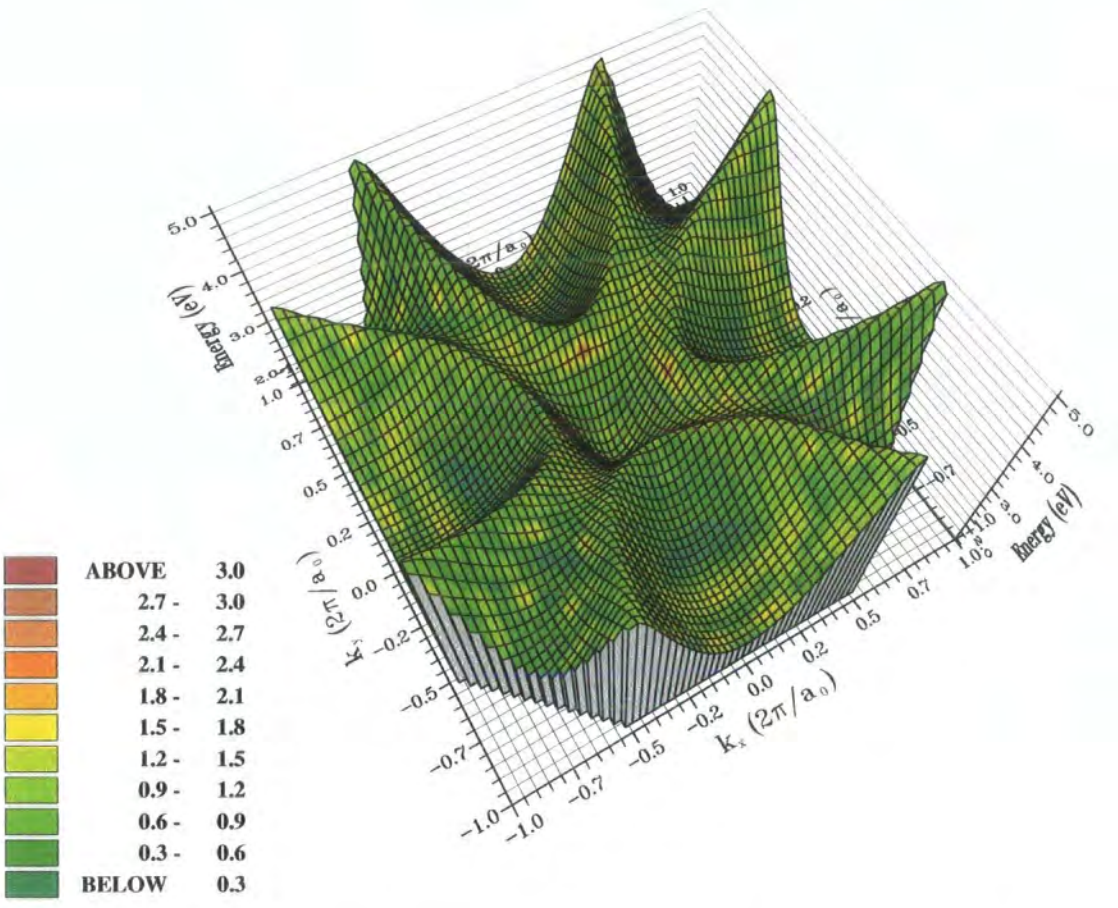


Figure 3.11: Interpolation errors on an adapted grid. The plot is the same type as that in Fig. 3.10. Note that the interpolation error is more-or-less uniform throughout the zone.

3.4 Wavefunction Interpolation

As discussed in §2.3.2, the pseudopotential calculation returns the wavefunctions in the form

$$\psi_b(\mathbf{r},\mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_b(\mathbf{r},\mathbf{k}) \tag{3.4}$$

where b is the band index, and the Bloch periodic part $u_b(\mathbf{r},\mathbf{k})$ is expressed as a sum of $2N$ plane waves[c]:

$$u_b(\mathbf{r},\mathbf{k}) = \frac{1}{\sqrt{\Omega}} \sum_{n=1}^{2N} \alpha_{b,n}(\mathbf{k}) e^{i\mathbf{G}_n\cdot\mathbf{r}} \tag{3.5}$$

where Ω is the volume of the crystal. Wavefunction data for the crystal is stored in the same way as energy data — on a grid of points distributed throughout the irreducible wedge. At each $\mathbf{k}$-point we must store the Bloch part $u_b(\mathbf{k})$, which means storing the coefficients $\alpha_{b,n}(\mathbf{k})$ evaluated at that point in $\mathbf{k}$-space. This requires storage of $2N$ complex numbers, or equivalently $4N$ real numbers. The storage requirements for the wavefunctions are thus $4N$ times greater than for the energies. In this work $N = 65$, and so storage of the wavefunction data places considerable demands on the computer memory.

3.4.1 Zone Centre Coefficients

The wavefunctions, returned by the pseudopotential calculation as expansions of plane waves, form an orthonormal set of basis functions. The plane wave basis would be complete if we used an infinite number of terms in the expansion. In practice we use a finite number N, which is chosen to provide sufficiently good convergence in quantities of interest such as transition matrix elements whilst not requiring unreasonable computational effort. (The time required to pre-calculate band structure information

[c]That is, N plane waves for the spin-up part of the wavefunction and N for the spin-down part. For the purposes of this chapter, it is not important that spin-up and spin-down terms exist, only that the expansion contains $2N$ terms.

at a point in **k**-space is $O(N^3)$).

The wavefunction can be expanded in any suitable basis set, and as will be discussed in this section, an alternative orthonormal basis set — the *zone centre wavefunctions* — exists that can accurately represent the wavefunction at most points in **k**-space using fewer terms than the plane wave basis.

The zone centre wavefunctions are the wavefunctions of the crystal evaluated at $\mathbf{k} = 0$. The zone centre wavefunction for the m^{th} band is $\phi_m(\mathbf{r})$, where

$$\phi_m(\mathbf{r}) \equiv \psi_m(\mathbf{r}, \mathbf{k} = 0) \equiv u_m(\mathbf{r}, \mathbf{k} = 0) \tag{3.6}$$

A Bloch periodic part of a wavefunction evaluated at some point in **k**-space can be expanded in terms of M zone centre wavefunctions:

$$u_b(\mathbf{r}, \mathbf{k}) = \sum_{m=1}^{M} \beta_{b,m}(\mathbf{k})\phi_m(\mathbf{r}) \tag{3.7}$$

where $\beta_{b,1} \ldots \beta_{b,M}$ are generally complex coefficients.

The pseudopotential calculation returns $2N$ bands of energy and wavefunction data (where N is the number of terms in the plane wave expansion). Thus there are $2N$ zone centre wavefunctions available for use as a basis set, i.e. in Eq. (3.7) $M \leq 2N$. If all $2N$ terms are used, the expansions of Eqs. (3.5) and (3.7) are exactly equal. If we use less than $2N$ terms in Eq. (3.7) then converting from a plane wave expansion to a zone centre expansion will 'lose' some of the wavefunction. This loss can be measured by l:

$$l = 1 - \left|\langle \psi_{pw} | \psi_{zc} \rangle\right|^2 \tag{3.8}$$

where ψ_{pw} and ψ_{zc} are the plane wave and zone centre expansions of some wavefunction. A value of $l = 0$ corresponds to the two representations being exactly equal, with l increasing as the accuracy of the zone centre expansion decreases.

We find that by using $M = 30$, the value of l is generally small ($l \lesssim 0.01$) for wavefunctions throughout most of the Brillouin zone. Thus by using the zone cen-

tre basis set instead of the plane wave set, we can expand the wavefunctions of the crystal in terms of just 30 complex coefficients instead of 130, without significant loss of accuracy. The saving in memory requirement of more than a factor of four is of particular importance as the large quantity of wavefunction data poses considerable storage problems.

Thus the interpolation scheme interpolates the coefficients $\beta_{b,1} \ldots \beta_{b,M}$ from a grid pre-calculated **k**-points. These zone centre coefficients can be used directly, for example in the case of simple overlaps $\langle \mathbf{k}_f | \mathbf{k}_i \rangle$, or converted back to the original plane wave expansion for evaluation of more general matrix elements $\langle \mathbf{k}_f | \widehat{O} | \mathbf{k}_i \rangle$ where $\widehat{O}$ is some operator.

The details of the transformation between plane wave and zone centre representations of the wavefunctions are set out in Appendix A.

Symmetry of the Zone Centre Wavefunctions

As with storage of the energy data, we can make use of the 48-fold symmetry of the Brillouin zone, thus reducing the volume in which wavefunction data is stored to the irreducible wedge. However, while energy for a given band at an arbitrary point in the Brillouin zone can be obtained straightforwardly from energy in the irreducible wedge, the corresponding operation is not so simple when dealing with wavefunction coefficients. The symmetry operations required to transform a point $\mathbf{k}_w$ in the irreducible wedge to a corresponding point $\mathbf{k}_a$ elsewhere in the zone must also be applied to the wavefunction itself. Thus, while the energy at $\mathbf{k}_a$ is simply the same as that at $\mathbf{k}_w$, this is clearly not so for the set of wavefunction coefficients at these two points.

The zone centre wavefunctions have symmetry properties which make the application of the operations in Table 3.2 a simpler matter than would be the case if they were applied to wavefunctions expanded as plane waves. This improves the speed at which wavefunction data can be obtained at general **k**-points from the interpolation scheme. Figs. 3.12 and 3.13 show the symmetry of the charge density associated with

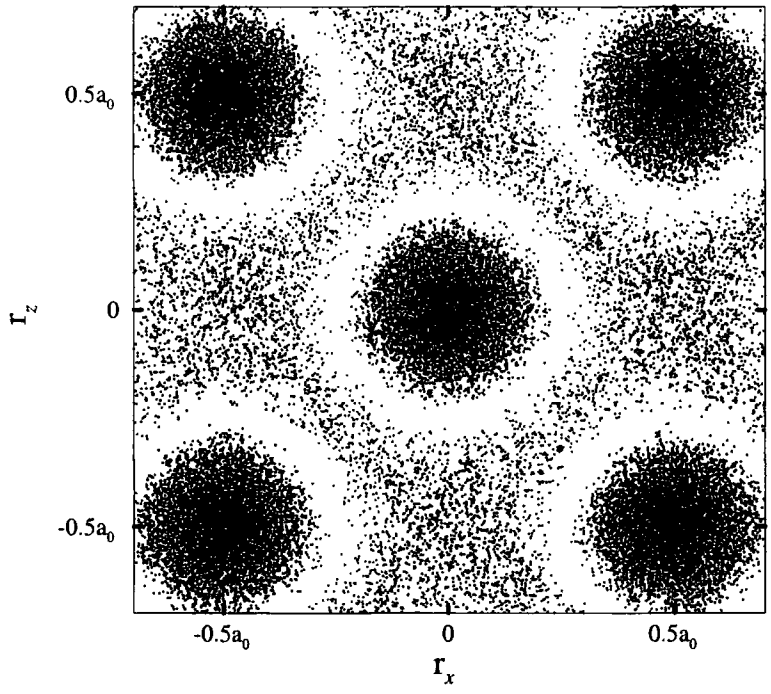


Figure 3.12: The probability density of the zone centre wavefunction for band 9 (the 1st conduction band) of GaAs, with symmetry $i|S\uparrow\rangle$. The density of dots denotes the electron probability density. Five anions (regions of high density) can be seen in the $\mathbf{r}_y = 0$ plane of the plot.

two example zone centre wavefunctions for GaAs.

3.4.2 Implementing the Interpolation Scheme

The real and imaginary parts of each zone centre coefficient used in the expansion of the wavefunction are interpolated in the same way as for energy data, i.e. quadratically within interpolating elements of the type shown in Fig. 3.5. As with the energy data, these elements are stored in the form of a grid filling the volume of the irreducible wedge and extending just outside it. Note that the amount of data to be handled is much greater — a single band of wavefunction data is equivalent to 60 bands worth of energy data, corresponding to the real and imaginary parts of the 30 expansion coefficients.

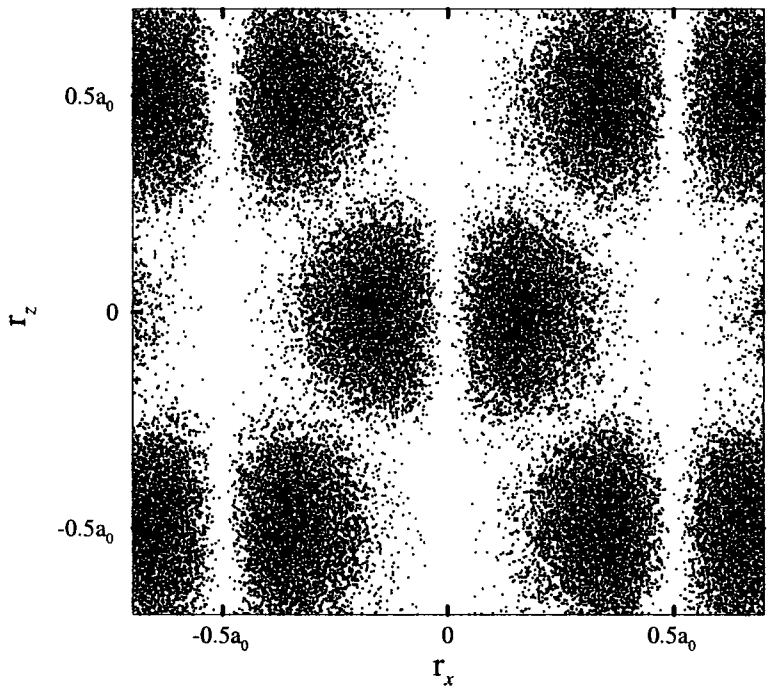


Figure 3.13: The probability density of the zone centre wavefunction for band 5 (the light hole band) of GaAs, with symmetry $\frac{1}{\sqrt{2}}|(X+iY)\uparrow\rangle$. As with Fig 3.12, the plot is in the $\mathbf{r}_y = 0$ plane, and so only the p_x-symmetry of the wavefunction is apparent.

The accuracy of the interpolated wavefunctions can be characterised by the value

$$\delta = 1 - \left|\langle\psi_i|\psi_c\rangle\right|^2. \tag{3.9}$$

where ψ_i and ψ_c are interpolated and calculated wavefunctions respectively. Perfectly interpolated wavefunctions correspond to $\delta = 0$, with δ increasing as the interpolation becomes worse.

It turns out that the wavefunction data is more difficult to interpolate accurately than the energy data. Fig. 3.14 indicates why. It shows the energy of band 9 (the first conduction band) plotted along the line L–Γ–X. Also plotted along the same line are the squared magnitudes of the 9^{th} and 13^{th} zone centre coefficients of the wavefunction. These coefficients correspond to zone centre wavefunctions with the same type of symmetry as those shown in Figs. 3.12 and 3.13 respectively. From the figure, it can be seen that the energy is a slowly varying function of $\mathbf{k}$, which is therefore easily interpolated. In contrast, the wavefunction coefficients vary rapidly at places along the line, particularly at $k \simeq (\frac{1}{3}00)$, where the character of the wavefunction changes from s-like to p-like. The rapid variation in certain regions of the irreducible wedge, which is typical of all the wavefunction coefficients, makes interpolation difficult and leads to large interpolation errors.

Another factor limiting the accuracy of the interpolated wavefunctions is the accuracy of the zone centre expansion. As explained in §3.4.1, this is generally high (within $\sim 1\%$). However regions of the irreducible wedge exist in which the value of l in Eq. (3.8) becomes significant for certain bands. The upper diagram of Fig. 3.15 shows the value of l plotted in the $k_z = 0$ plane of the irreducible wedge for band 5. In the region near K, l increases rapidly as the zone centre basis set, limited to 30 coefficients, fails to give an acceptable representation of the wavefunction. In such regions, the interpolation scheme cannot be used and wavefunction data must be obtained by direct application of the pseudopotential calculation. The lower diagram of Fig. 3.15 shows the intersection of the band 5 interpolation grid with the $k_z = 0$ plane

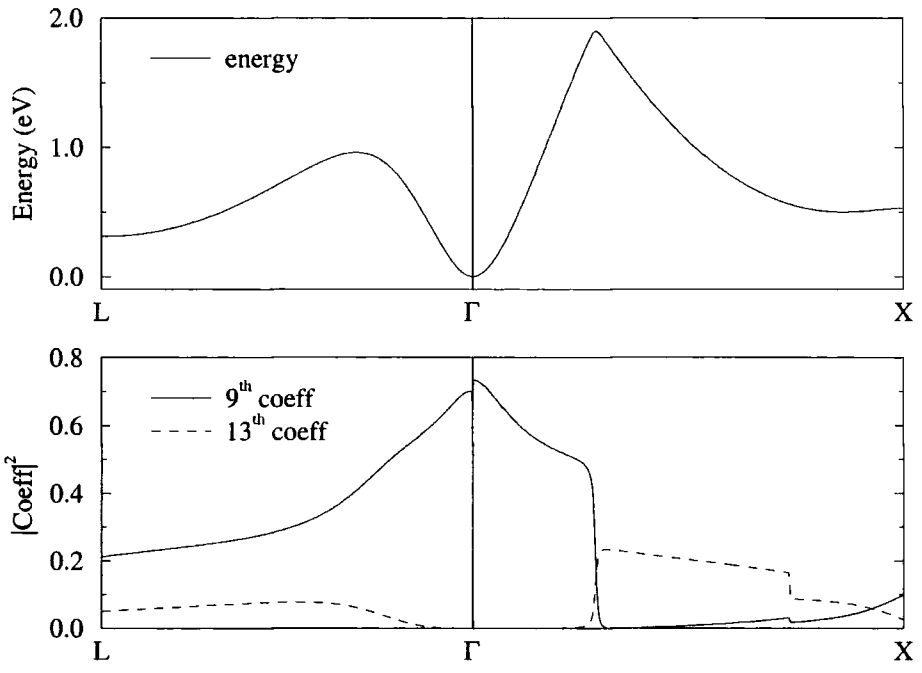


Figure 3.14: Band 9 (1st conduction band) energy and wavefunction data for GaAs, plotted along the line L–Γ–X. While the energy varies relatively slowly as a function of $\mathbf{k}$, the coefficients vary rapidly, particularly at $\mathbf{k} \simeq (\frac{1}{3}00)$.

of the irreducible wedge. No interpolating elements are defined throughout the region in which the zone centre representation of the wavefunction is poor.

Fig. 3.16 represents the overall algorithm for interpolating, or calculating where necessary, wavefunction data.

3.4.3 The Use of Adaptive Grids

Adaptive grids are not found to be effective in the interpolation of wavefunction data in the way that they are for energy data. In the case of energy interpolation, the criteria for adapting a region of the mesh are clear: if the error in the interpolated energy is above some threshold, the mesh is made finer at that point; if it is below some other threshold, the mesh is made coarser at that point.

In the case of the wavefunction interpolation, it is not possible to specify easily the criteria to define the interpolation error. The value of δ defined in Eq. (3.9) gives an indication of the accuracy of the interpolated wavefunction. However, in an application the quantity of interest is not the wavefunction itself but matrix elements obtained from sets of wavefunctions — sets of four wavefunctions, in the case of impact ionisation. The error on the value of such a matrix element calculated using interpolated data is not simply obtained from the interpolation errors of each of the individual wavefunctions.

Because it is not known *a priori* which matrix elements will be required in an application, there is no way of effectively adapting the wavefunction grids to ensure a uniformly low interpolation error in the matrix elements. Wavefunctions are therefore interpolated on regular grids. The only form of band dependent adaptation performed is the removal of regions in which the zone centre expansion fails, as discussed in §3.4.2.

3.4.4 Quality of the Interpolation

As with the energy eigenvalues, the performance of the wavefunction interpolation scheme is measured in terms of its memory efficiency, speed and accuracy.

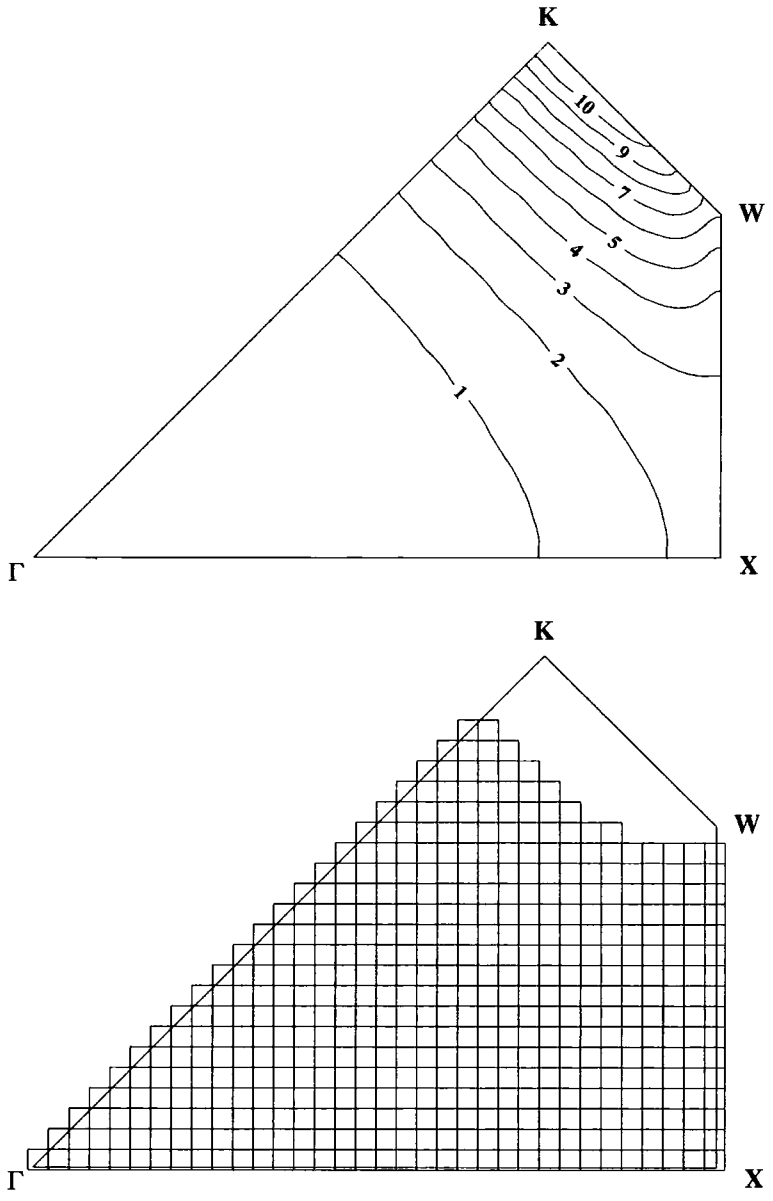


Figure 3.15: Loss of wavefunction accuracy in band 5 (light hole band) of GaAs due to an incomplete zone centre basis set. The upper diagram shows contours of l, as defined in Eq. (3.8), expressed as a percentage. The lower diagram shows the wavefunction interpolation grid, which is undefined in the region of high inaccuracy near K–W.

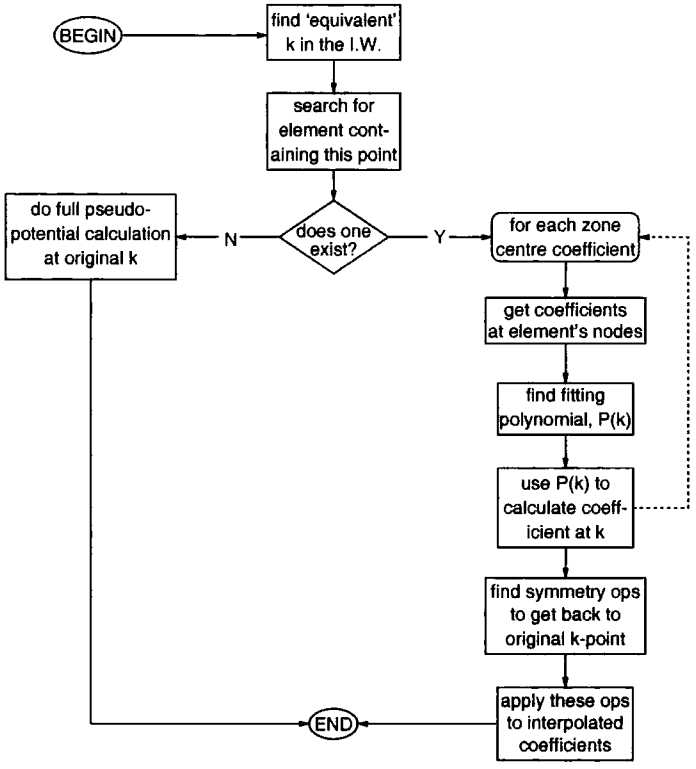


Figure 3.16: The algorithm for interpolation of wavefunction data, which resorts to the full pseudopotential calculation in uninterpolated regions of the wedge.

Memory

For each **k**-point at which wavefunction data is required, 30 complex zone centre coefficients must be stored — 60 single precision real numbers, each requiring four bytes of storage space. The wavefunction grids consist of about 34000 points for each band, of which some are removed from regions of the irreducible wedge in which the zone centre expansion for the given band's wavefunctions fails. The wavefunction data for 10 bands of GaAs can be stored in under 90 MBytes of RAM, which is available on many modern workstations.

Table 3.5 gives the volume of the irreducible wedge removed in each band, expressed as a percentage of the whole irreducible wedge.

Band	Vol. uninterpolated (%)
3,4	0.0
5,6	3.3
7,8	0.2
9,10	0.5
11,12	5.0

Table 3.5: The percentage of the volume of the irreducible wedge which, due to poor zone centre representation of the wavefunction, is left uninterpolated in each band for GaAs.

Speed

As with the energy interpolation scheme, a Hewlett-Packard 735 workstation was used to compare the speed of interpolating and calculating the wavefunction data. It was found that the direct pseudopotential calculation of the wavefunction could be performed 0.65 times per second, while 1100 interpolations could be carried out each second — an increase in speed by a factor of 1800.

The overall rate at which wavefunction information can be obtained from the interpolation scheme is less than 1100 **k**-points per second, due to the need to perform the full calculation in uninterpolated regions corresponding to poor zone centre expan-

sion. The exact reduction in the rate depends on the fraction of points required during execution of the application which lie in such regions. Fortunately it turns out that impact ionisation transitions involve states in the uninterpolated regions only rarely. This is because the zone centre expansion tends to fail for states corresponding to the highest carrier energies (i.e. the higher electron energies in the conduction bands and lower electron energies in the valence bands). The states for which band structure data is required throughout the zone — the generated hole and final state electrons — generally lie at lower energies, and hence in regions for which the interpolation grid is defined. Therefore the total interpolation rate is less than 1100 **k**-vectors per second, but not considerably so.

Accuracy

The accuracy of the wavefunction interpolation is tested in the same way as for the energy interpolation — by comparing the interpolated and calculated wavefunctions for a large number of **k**-points picked at random throughout the irreducible wedge. Table 3.6 shows the accuracy obtained from the interpolated wavefunctions. The accuracy is given in terms of δ, defined in Eq. (3.9) and expressed as a percentage. The smaller the value of δ, the better the interpolation, with $\delta = 0$ corresponding to perfect interpolation. Any error introduced due to imperfect representation of the wavefunction using the zone centre basis set is also included in these results.

Band	RMS value of δ (%)
3,4	0.7
5,6	3.2
7,8	2.2
9,10	2.9
11,12	3.8

Table 3.6: Accuracy of the wavefunction interpolation, averaged over a large number of **k**-points chosen randomly throughout the zone. The parameter δ is defined in Eq. (3.9).

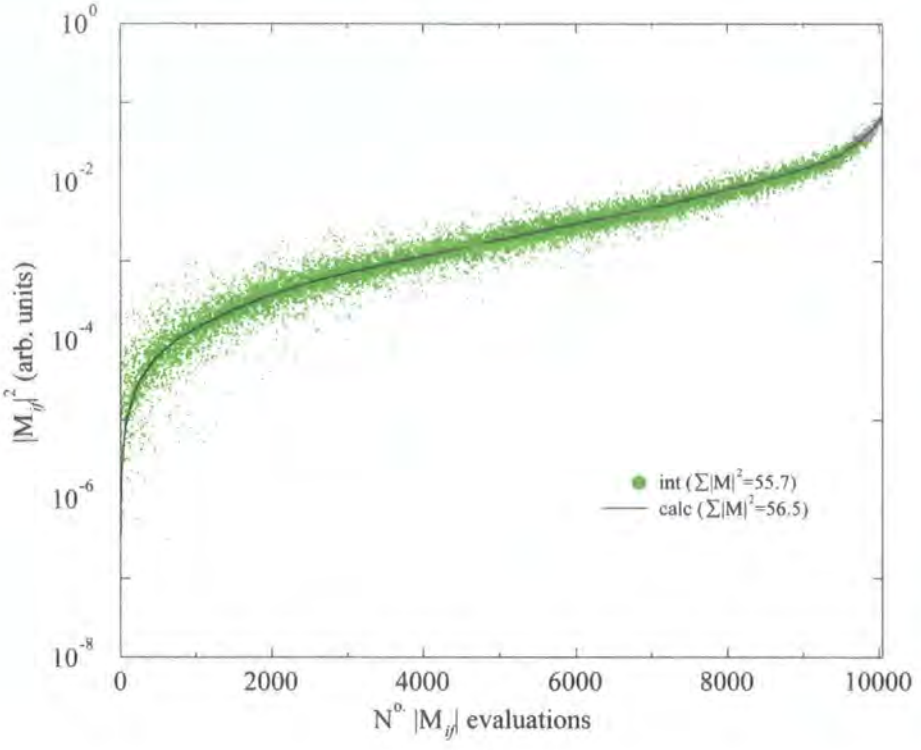


Figure 3.17: Comparison of impact ionisation matrix elements obtained from interpolated and calculated wavefunctions (sorted into order of increasing magnitude of calculated element). The elements all correspond to energy and momentum conserving transitions initiated by electrons in the 1^{st} conduction band of $In_{0.53}Ga_{0.47}As$. Although interpolation errors on individual elements are often large, the total obtained by interpolation is within $\sim$ 2% of that obtained by calculation. Similar accuracy is obtained for transitions initiated from the second conduction band

As explained in §3.4.3, the value of δ only acts as a guide to the accuracy of the interpolation, and the quantities of interest are actually matrix elements. Fig. 3.17 compares impact ionisation matrix elements obtained from wavefunction data produced by the direct application of the pseudopotential calculation, and via the interpolation scheme. From the figure, it is clear that individual matrix elements M_i obtained from interpolated wavefunction data are generally poor approximations to the equivalent matrix elements M_c obtained from calculated data. However, the trend in the values of M_i follows the variation of M_c, and the value obtained by summing all M_i's is a good approximation to the equivalent value for the M_c's. Since the rate integration involves summing matrix elements in this way, the overall accuracy of the rate obtained from

interpolated wavefunctions is high.

3.5 Epsilon Interpolation

The dielectric function of the crystal $\epsilon(\mathbf{q},\omega)$ appears in the expression for the impact ionisation matrix element (as discussed in Chapter 4), and so we must be able to obtain values for ϵ as a function of $\mathbf{q}$ and ω rapidly. The calculation of ϵ is computer intensive and so, as with energies and wavefunctions, values for it are pre-calculated and stored on a grid of points. During running of the application, values of ϵ at general $(\mathbf{q},\omega)$ are interpolated from the values in the grid.

The pre-calculation is performed using the expressions for the real and imaginary parts of the dielectric function given in Eqs. (2.33) and (2.34) of Chapter 2. These are combined in the expression

$$\epsilon(\mathbf{q},\omega) = 1 + \frac{e^2}{\Omega_C \epsilon_0 q^2} \sum_{\mathbf{k}} f(\mathbf{k},\mathbf{q},\omega) \tag{3.10}$$

where Ω_C is the volume of the crystal and the function $f(\mathbf{k},\mathbf{q},\omega)$ is given by

$$f(\mathbf{k},\mathbf{q},\omega) = \sum_{c,v} \left| \langle u^c_{\mathbf{k}} \left| u^v_{\mathbf{k}+\mathbf{q}} \right\rangle \right|^2 \times \left\{ \left[\frac{1}{E_{cv} - \hbar\omega} + \frac{1}{E_{cv} + \hbar\omega} \right] + i \left[\pi\delta(E_{cv} - \hbar\omega) \right] \right\} \tag{3.11}$$

in which $E_{cv} = E_c(\mathbf{k}) - E_v(\mathbf{k}+\mathbf{q})$, and the remaining symbols have the same meanings as in Eqs. (2.33) and (2.34). The expression of Eq. (3.10) is re-written as

$$\epsilon(\mathbf{q},\omega) = 1 + \frac{e^2 \Omega_{BZ}}{(2\pi)^3 \epsilon_0 q^2} \overline{F}(\mathbf{q},\omega) \tag{3.12}$$

where Ω_{BZ} is the volume of the Brillouin zone and $\overline{F}(\mathbf{q},\omega)$ is the average value of $f(\mathbf{k},\mathbf{q},\omega)$ throughout Ω_{BZ}. The problem then is to determine the value of $\overline{F}(\mathbf{q},\omega)$, and this is done numerically by a Monte Carlo method as follows.

The value of $\overline{F}(\mathbf{q},\omega)$ is determined at fixed $\mathbf{q}$. A value of $\mathbf{k}$ is picked at random in the Brillouin zone and the energies and wavefunctions in each band at $\mathbf{k}$ and $\mathbf{k}+\mathbf{q}$

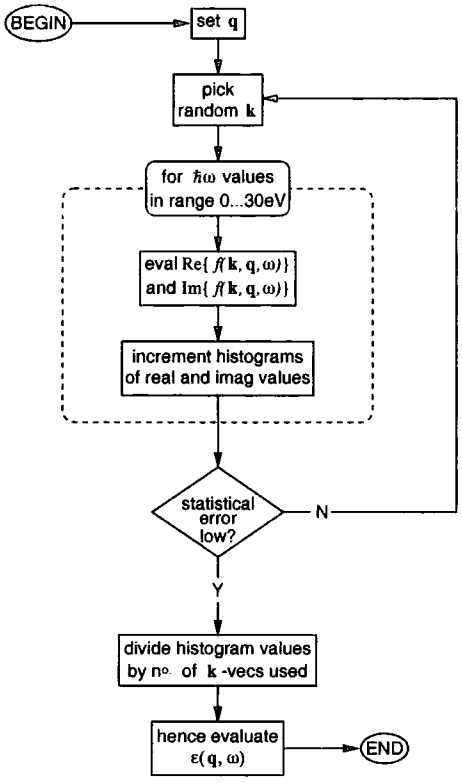


Figure 3.18: The Monte Carlo algorithm to integrate real and imaginary parts of the dielectric function at a given **q**-vector. The algorithm must be run several times to get $\epsilon(\mathbf{q},\omega)$ throughout the range of **q**-values of interest.

calculated. The real and imaginary parts of $f(\mathbf{k},\mathbf{q},\omega)$ at the given **k** and **q** are then calculated at $\hbar\omega$ values ranging in small steps from 0–30eV and stored in histograms with respect to energy (one histogram for each of the real and imaginary parts). This procedure is repeated at fixed **q** for a large number of random **k** vectors, each time adding to the histograms of stored real and imaginary parts. After a large number of such evaluations, the values in the histograms are divided by the number of **k**-points sampled to give the real and imaginary parts of $\overline{F}(\mathbf{q},\omega)$ at given **q** and a range of ω. Hence the real and imaginary parts of ϵ as a function of ω at given **q** are obtained. The algorithm is summarised in Fig. 3.18

3.5.1 Approximations in the Numerical Integration

The evaluation of $f(\mathbf{k},\mathbf{q},\omega)$ over the energy range of $\hbar\omega$ requires certain approximations to avoid numerical difficulties.

In the evaluation of the real part of $f(\mathbf{k},\mathbf{q},\omega)$ over the energy range of $\hbar\omega$ values, care must be taken to ensure that the expression in the first set of brackets in Eq. (3.11)

does not become very large as $\hbar\omega \rightarrow E_{cv}$, as this leads to spikes appearing in the values of $\epsilon(\mathbf{q},\omega)$. To avoid this error, the term in brackets is approximated by

$$\frac{1}{E_{cv}-\hbar\omega}+\frac{1}{E_{cv}+\hbar\omega} \simeq \mathcal{R}e\left\{\frac{1}{E_{cv}-\hbar\omega-i\eta}+\frac{1}{E_{cv}+\hbar\omega+i\eta}\right\} \tag{3.13}$$

where η is a small positive value. When $\hbar\omega \not\simeq E_{cv}$, the right hand side of Eq. (3.13) is a good approximation to the left hand side. As $\hbar\omega \rightarrow E_{cv}$ and the left hand side becomes very large, the right hand side tends to zero. Thus the spikes in $\epsilon(\mathbf{q},\omega)$ are avoided.

In the evaluation of the imaginary part of $f(\mathbf{k},\mathbf{q},\omega)$, the Dirac delta function $\delta(E)$ in the second pair of brackets must be approximated by some function of finite width and unit area. In this case, a top-hat function $h(E)$ is used:

$$\delta(E_{cv}-\hbar\omega) \simeq h(E_{cv}-\hbar\omega) = \begin{cases} \frac{1}{2\delta e} & \text{if } |E_{cv}-\hbar\omega| \leq \delta e, \\ 0 & \text{otherwise.} \end{cases} \tag{3.14}$$

where δe is a small energy value.

Values for η and δe are chosen to ensure that $\epsilon(\mathbf{q},\omega)$ has converged with respect to them, i.e. that the dielectric function does not change significantly with small changes in η and δe. The numerical accuracy of the integrals can also be checked by comparing the values of the real and imaginary parts of $\epsilon(\mathbf{q},\omega)$ obtained by direct calculation with those obtained by application of the Kramers-Kronig relations (see §2.4). This comparison is made for GaAs in Fig. 2.5 of Chapter 2.

3.5.2 Isotropic $\epsilon(\mathbf{q},\omega)$ Approximation

To obtain values of $\epsilon(\mathbf{q},\omega)$ at positions throughout $\mathbf{q}$-space as well as ω-space, the numerical evaluation of $\overline{F}(\mathbf{q},\omega)$ and hence $\epsilon(\mathbf{q},\omega)$ is performed separately for different $\mathbf{q}$-vectors. In principle, $\epsilon(\mathbf{q},\omega)$ would be interpolated as a function of four variables, corresponding to its variation with respect to the three components of $\mathbf{q}$ and to ω. However, the number of mesh points in the 4-dimensional interpolation grid required

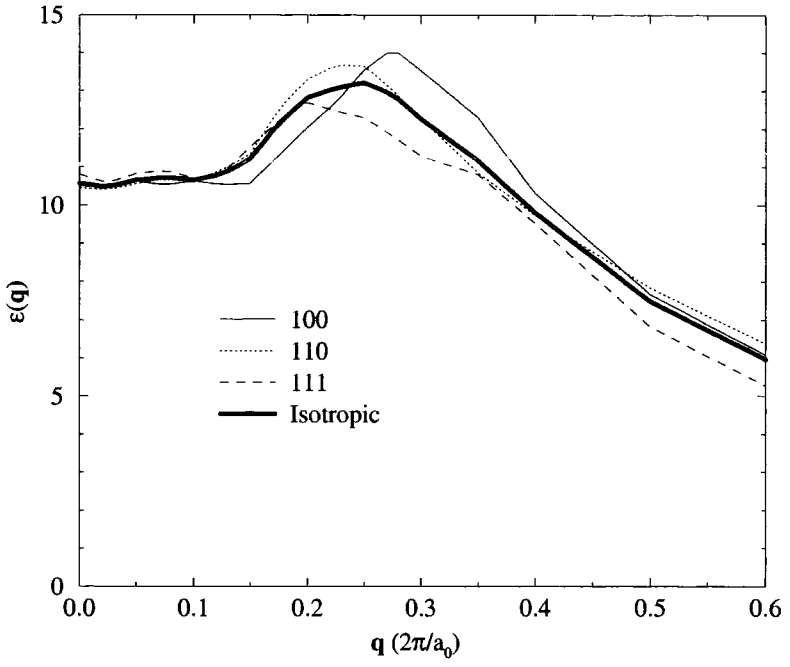


Figure 3.19: The anisotropic dielectric function of GaAs plotted in the [100], [110] and [111] directions, and its isotropic approximation. Each line is plotted at $\omega = 4$eV.

for such a function would be very large, leading to problems both in the time required to pre-calculate the dielectric function data, and the memory required to store the grid. Therefore the full **q**-dependent expression for $\epsilon(\mathbf{q}, \omega)$ is replaced with an isotropic approximation $\epsilon(q, \omega)$, defined as:

$$\epsilon(q,\omega) = \frac{1}{26}\Big[6 \times \epsilon(q_{100},\omega) + 12 \times \epsilon(q_{110},\omega) + 8 \times \epsilon(q_{111},\omega)\Big] \qquad (3.15)$$

where $\epsilon(q_{abc}, \omega)$ is $\epsilon(q, \omega)$ evaluated along the [abc]-direction. Fig. 3.19 compares in the case of GaAs the true anisotropic form of $\epsilon(\mathbf{q}, \omega)$ for certain directions of **q** with the isotropic approximation, $\epsilon(q, \omega)$. It can be seen that the variation of $\epsilon(\mathbf{q}, \omega)$with direction is not great, and the full anisotropic function is well approximated by an isotropic one.

Use of the isotropic expression for ϵ reduces it to a function of just two variables —

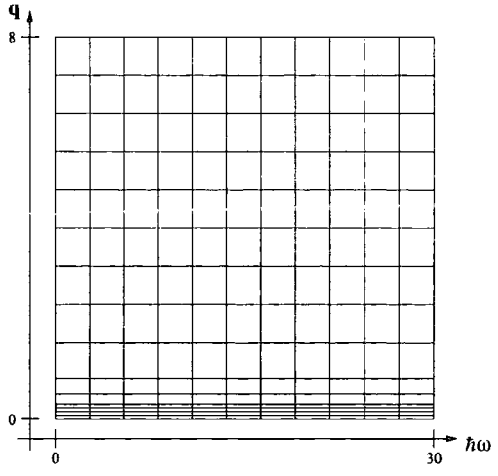


Figure 3.20: Schematic representation of the interpolation grid used for $\epsilon(q,\omega)$. The actual grid is denser than the one represented here.

q and ω — and so the interpolation grid is 2-dimensional. Fig. 3.20 shows the form of the grid. Values of ϵ are calculated and stored at the (q,ω) coordinates corresponding to the intersection of the lines. Within the rectangular elements, ϵ is interpolated bilinearly. Note that as well as being only 2-dimensional, it is a simpler form of grid to that used for energy band structure interpolation. Fig. 3.21 shows ϵ plotted as a function of q and ω.

3.5.3 Use of Calculated Band Structure

In the evaluation of $\epsilon(\mathbf{q},\omega)$, the interpolation scheme does not give good results as $q \rightarrow 0$. In this limit the magnitude of the matrix element $\langle u^c_{\mathbf{k}} | u^v_{\mathbf{k}+\mathbf{q}} \rangle$ also tends to zero due to the orthogonality of the wavefunctions in different bands but at the same $\mathbf{k}$. The interpolated wavefunctions are only approximately orthogonal, and so their matrix elements do not tend exactly to zero as $\mathbf{q} \rightarrow 0$. Thus the *absolute* interpolation error remains finite and the *percentage* error becomes very large, which leads to correspondingly large percentage errors in the value of $\epsilon(\mathbf{q},\omega)$. The dielectric calculation at small $\mathbf{q}$ is therefore carried out by direct application of the full pseudopotential calculation, which is computer intensive but gives correctly orthogonalised wavefunctions.

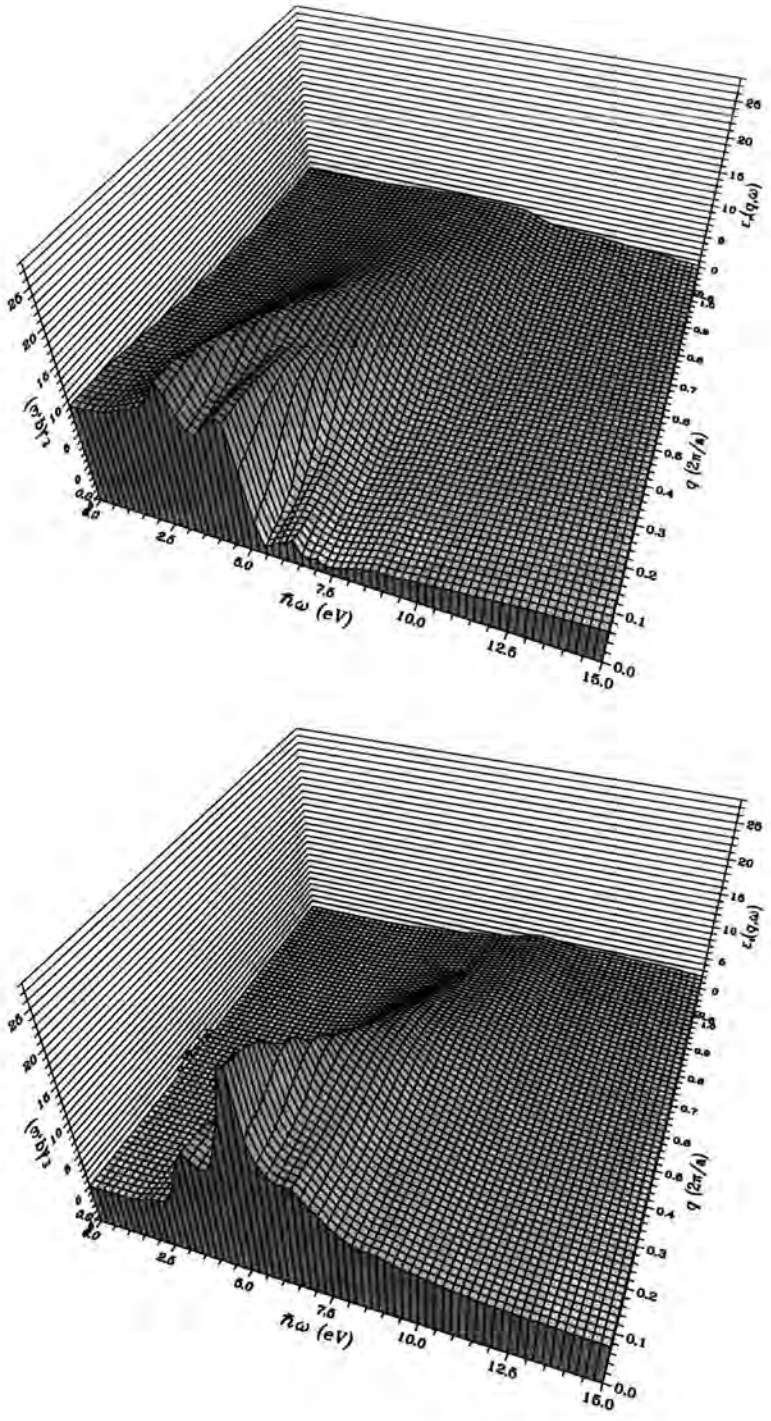


Figure 3.21: The dielectric function of GaAs, as a function of q and ω. The upper plot shows the real part, the lower plot the imaginary part.

This is an illustration of the point made in §3.4.3: limiting the interpolation error on individual wavefunctions, measured in terms of δ defined in Eq. (3.9), does not necessarily limit the interpolation error present in quantities of interest, i.e matrix elements.

The implications of this for the calculation of impact ionisation rates are not severe, however. The wavefunction overlap integrals involved in the calculation of impact ionisation matrix elements also tend to zero as $\mathbf{q} \to 0$ (see Chapter 4). However in a rate calculation the majority of transitions correspond to finite $\mathbf{q}$, and the problem of lack of orthogonality of interpolated wavefunctions does not affect the result significantly.

Chapter 4

Impact Ionisation: Theory

Band-to-band impact ionisation is the process in which a high energy carrier excites an electron from the valence band to the conduction band, thus creating two new charge carriers (see, for example, [14], [100]). One such process is shown in Fig. 4.1. It shows a high energy conduction band electron, labelled with its wavevector $\mathbf{k}_1$ and known as the *impacting* electron, being scattered by a valence band electron labelled $\mathbf{k}_2$ and known as the *impacted* electron. The result of the process is that the impacted electron is excited from the valence band to a state near the bottom of the conduction band, the energy to achieve this being supplied by the impacting electron, which is also scattered to a state near the bottom of the conduction band. These two final states are labelled $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$. Thus, where before the process takes place there is one carrier (the impacting electron), after there are three (the two conduction band electrons and the hole in the valence band).

This is an example of *electron initiated* impact ionisation, and many similar processes can take place involving various combinations of bands. In each case, an electron is excited from the valence to the conduction band, resulting in the creation of two new charge carriers — an electron and a hole. Fig. 4.2 shows two other examples of electron initiated processes.

Care must be taken in considering the generated hole. The process of impact

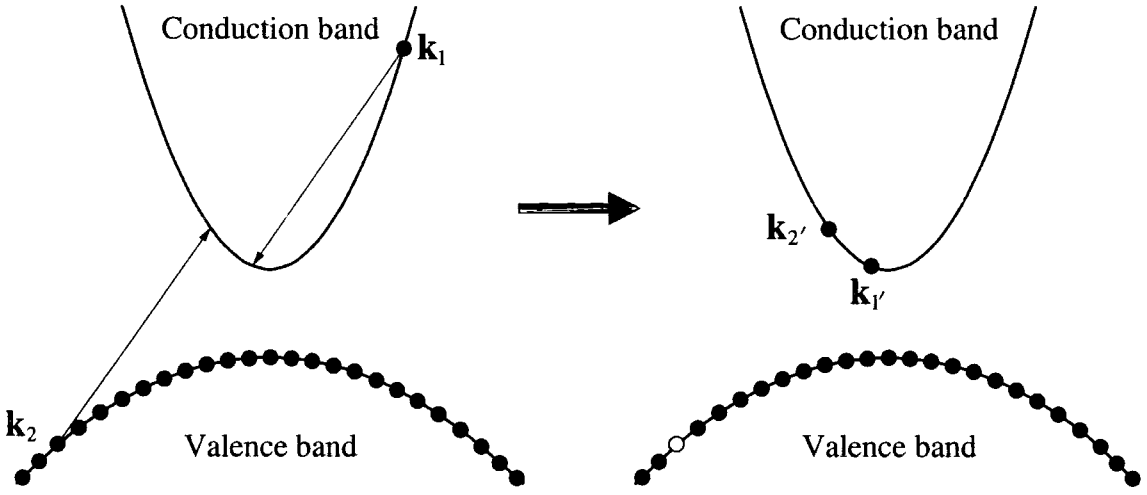


Figure 4.1: A schematic representation of an impact ionisation process. The left-hand side shows the electrons in their initial states $\mathbf{k}_1$ and $\mathbf{k}_2$. There is only one charge carrier – the electron at $\mathbf{k}_1$. On the right-hand side, the electrons are in their final states, $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$. Now there are three charge carriers – the electrons $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$ and the hole associated with the vacancy remaining at $\mathbf{k}_2$.

ionisation leaves an unoccupied state at $\mathbf{k}_2$ in the valence band. This corresponds to a positively charged carrier, i.e. a hole, at $-\mathbf{k}_2$. Thus the generated hole lies at minus the wavevector of the impacted electron [101].

Impact ionisation can also be initiated by high energy holes, a process which can be thought of as the exact analogue of the electron initiated process. The example in Fig 4.3 shows a high energy hole in the valence band, $\mathbf{k}_1$, which excites a hole in the conduction band, $\mathbf{k}_2$. The final state consists of two holes, $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$, at low energy in the valence band and an electron remaining in the conduction band.

The hole initiated process can alternatively be considered in terms of transitions made by electrons. Fig 4.4 shows exactly the same process as Fig 4.3, but this time represented as a change in occupancy of electron states. Here the process is initiated by an electron at the top of the valence band, $\mathbf{k}_{1'}$, dropping down to fill a vacant state, $\mathbf{k}_1$, at lower energy. The energy made available is taken up by a valence band electron $\mathbf{k}_{2'}$, which is excited to an unoccupied conduction band state, $\mathbf{k}_2$.

The reverse of impact ionisation is *Auger recombination* [100], which provides a mech-

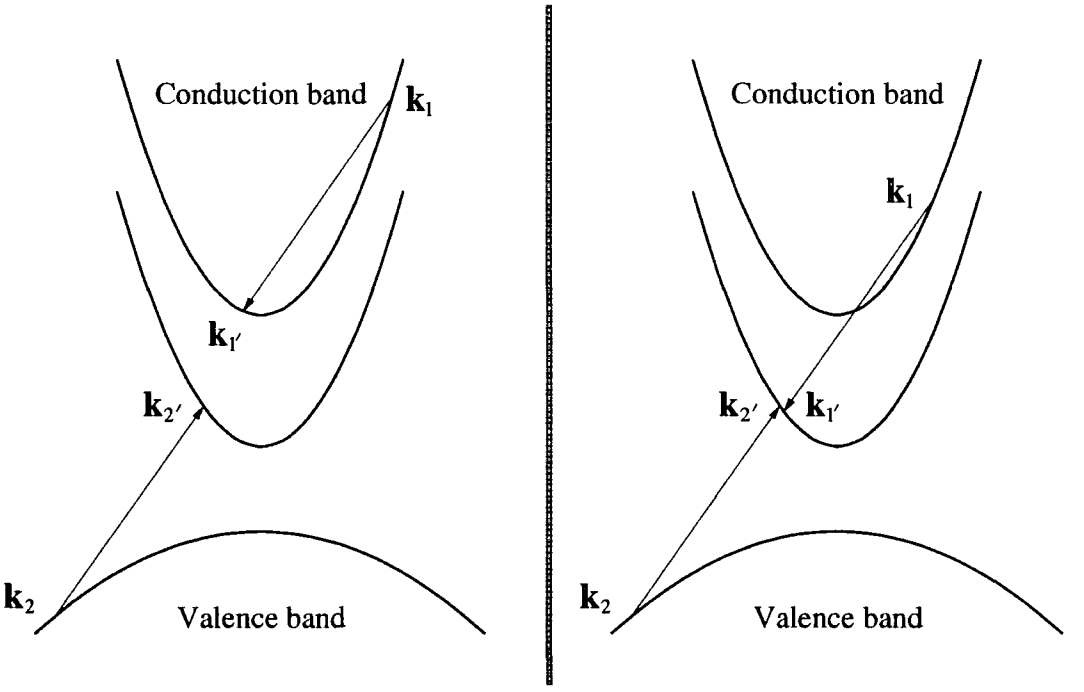


Figure 4.2: Schematic representations of other possible transitions, involving higher conduction bands. The impacting and impacted electrons need not have final states in the same band (as in the left-hand diagram), nor need the impacting carrier remain in the same conduction band (as in the right-hand diagram). In principle, up to four different bands may be involved.

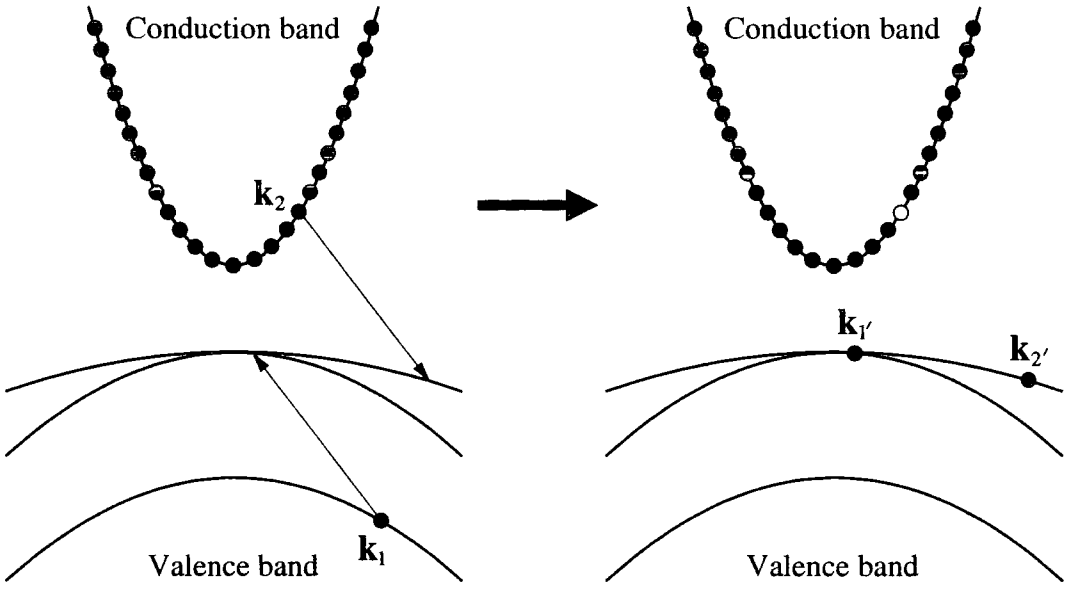


Figure 4.3: Hole initiated impact ionisation. A high energy hole in the valence band (hole energy increases *down* the y-axis) ionises a hole in the conduction band (which is normally full of holes), leaving behind an electron. The event represented here is identical to that represented in Fig. 4.4.

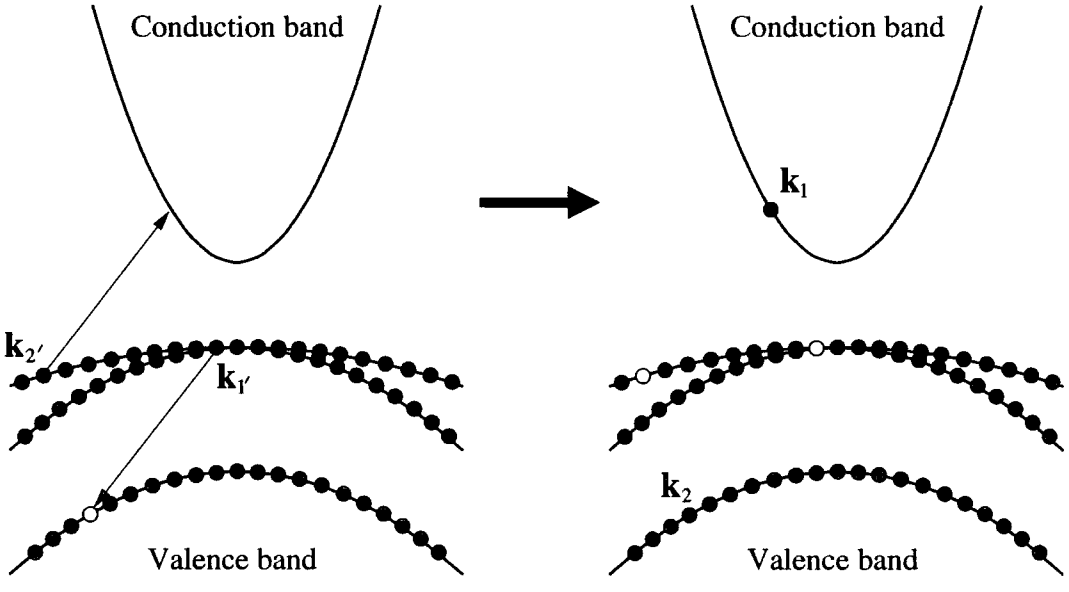


Figure 4.4: Hole initiated impact ionisation. This event is identical to that represented in Fig. 4.3. Here, it is shown in terms of transitions made by electrons instead of holes. The electron at $\mathbf{k}_{1'}$ initiates the process, losing energy and promoting the electron at $\mathbf{k}_{2'}$ to the conduction band as a result.

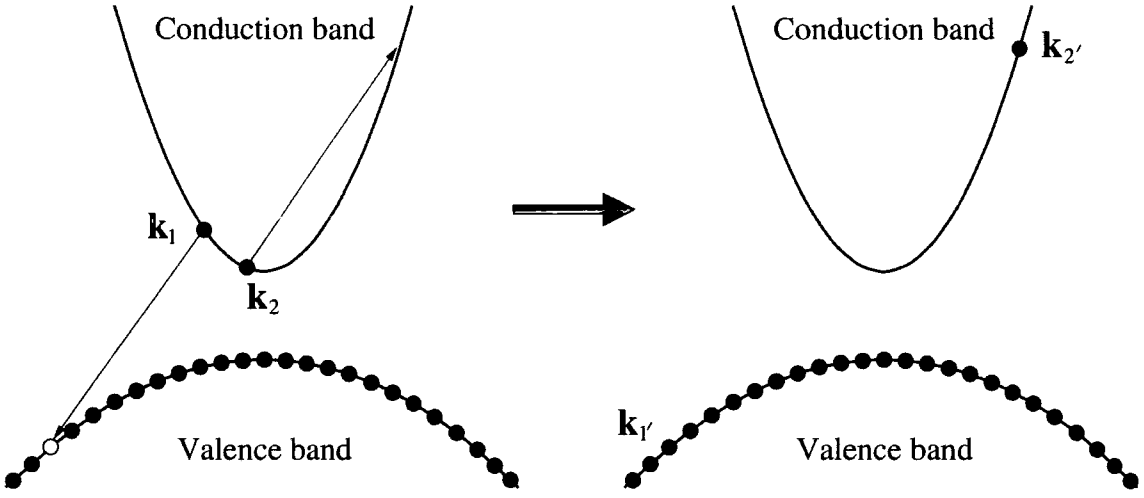


Figure 4.5: A schematic representation of an Auger recombination process. On the left is the initial state with three charge carriers — the electrons at $\mathbf{k}_1$ and $\mathbf{k}_2$ and the hole associated with the vacancy at $\mathbf{k}_{1'}$. On the right is the final state with just one carrier at $\mathbf{k}_{2'}$. The electron-hole pair $\mathbf{k}_1$-$\mathbf{k}_{1'}$ have recombined. Note that this figure is the same as Fig. 4.1, but with time reversed.

anism for the non-radiative recombination of electron-hole pairs. In this process, a low energy electron in the conduction band makes a transition to a vacant state in the valence band. The energy made available excites another conduction band electron, or a valence band hole, to a higher energy. One such process is represented in Fig. 4.5. Note that this figure depicts the same transition as Fig. 4.1, but with time reversed.

As with other carrier scattering mechanisms, when considered in the Fermi's Golden Rule approximation energy and crystal momentum conservation apply to the impact ionisation and Auger recombination process, i.e.

$$E_{1'} + E_{2'} = E_1 + E_2 \tag{4.1}$$

$$\mathbf{k}_{1'} + \mathbf{k}_{2'} = \mathbf{k}_1 + \mathbf{k}_2 + \mathbf{G} \tag{4.2}$$

where E_1, E_2, $E_{1'}$ and $E_{2'}$ are the energies of the states at $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$ respectively, and $\mathbf{G}$ is a reciprocal lattice vector.

The processes described above do not involve interaction with phonons. However, processes can occur in which one or more phonons are created or annihilated, and for

such 'phonon assisted' processes, the conservation laws of Eqs. (4.1) and (4.2) must be amended to include the energy and wavevector of the phonon(s) involved. Since phonon assisted processes are a second order effect, they are neglected in this work, and only scattering rates for which Eqs.(4.1) and (4.2) are applicable are considered. However, future study may show that the phonon assisted transitions can become important near threshold, where the first order transitions are highly restricted by energy and momentum conservation.

To determine the total rate at which an initiating carrier at $\mathbf{k}_1$ in a given band will be scattered via the process of impact ionisation or Auger recombination, the rates due to all distinct transitions to all possible final states must be summed over.

For brevity, only electron initiated impact ionisation processes will be considered explicitly in the remainder of this chapter. However, it is straightforward to adapt the theory discussed to the case of hole initiated impact ionisation and to electron or hole initiated Auger recombination.

4.1 The Transition Rate

The rate of transition due to impact ionisation for two electrons initially in states at $\mathbf{k}_1$ and $\mathbf{k}_2$[a] to final states at $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$ is given by Fermi's Golden Rule [100]:

$$R_{II}(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_{1'}, \mathbf{k}_{2'}) = \frac{2\pi}{\hbar} |M_{if}|^2 \, \delta(E_{1'} + E_{2'} - E_1 - E_2) \tag{4.3}$$

where the energy conservation expressed in Eq. (4.1) is ensured through the Dirac delta function, and as will be discussed in the next section, the crystal momentum conservation of Eq. (4.2) is ensured by the matrix element.

To obtain the total rate of scattering due to all possible transitions that a given impacting carrier at $\mathbf{k}_1$ can undergo, we must integrate Eq. (4.3) over all final states[b] for which the Dirac delta function is non-zero. This will involve evaluating the matrix

[a] To simplify the notation, $\mathbf{k}$ is used here to denote a position in $\mathbf{k}$-space and a band index.

[b] Care must be taken to include each distinct transition only once — see §5.5.

element M_{if} at points throughout the Brillouin zone.

4.2 The Matrix Element

The impact ionisation matrix element is of the familiar form

$$M_{if} = \int \Psi_f^* V \Psi_i \, d\tau \tag{4.4}$$

where Ψ_i and Ψ_f are the wavefunctions of the initial and final states respectively, and V is the appropriate operator representing the perturbation that causes the process.

The situation for impact ionisation is complicated by the fact that two electrons take part, and so the initial and final states must have the required anti-symmetry, that is [100,102]

$$\Psi_i = \frac{1}{\sqrt{2}}\Big[\psi_1(\mathbf{r_1})\psi_2(\mathbf{r_2}) - \psi_2(\mathbf{r_1})\psi_1(\mathbf{r_2})\Big] \tag{4.5}$$

$$\Psi_f = \frac{1}{\sqrt{2}}\Big[\psi_{1'}(\mathbf{r_1})\psi_{2'}(\mathbf{r_2}) - \psi_{2'}(\mathbf{r_1})\psi_{1'}(\mathbf{r_2})\Big] \tag{4.6}$$

where spin has been neglected for notational simplicity. Substituting the wavefunctions of Eqs. (4.5) and (4.6) in Eq. (4.4) gives

$$M_{if} = \frac{1}{2}\int \Big[\psi_{1'}^*(\mathbf{r_1})\psi_{2'}^*(\mathbf{r_2}) - \psi_{2'}^*(\mathbf{r_1})\psi_{1'}^*(\mathbf{r_2})\Big] V \Big[\psi_1(\mathbf{r_1})\psi_2(\mathbf{r_2}) - \psi_2(\mathbf{r_1})\psi_1(\mathbf{r_2})\Big] d^3\mathbf{r}_1 d^3\mathbf{r}_2 \tag{4.7}$$

which, multiplied out gives

$$\begin{aligned} M_{if} = \frac{1}{2}\int \Big[& \psi_{1'}^*(\mathbf{r_1})\psi_{2'}^*(\mathbf{r_2}) V \psi_1(\mathbf{r_1})\psi_2(\mathbf{r_2}) \\ & + \psi_{2'}^*(\mathbf{r_1})\psi_{1'}^*(\mathbf{r_2}) V \psi_2(\mathbf{r_1})\psi_1(\mathbf{r_2}) \\ & - \psi_{1'}^*(\mathbf{r_1})\psi_{2'}^*(\mathbf{r_2}) V \psi_2(\mathbf{r_1})\psi_1(\mathbf{r_2}) \\ & - \psi_{2'}^*(\mathbf{r_1})\psi_{1'}^*(\mathbf{r_2}) V \psi_1(\mathbf{r_1})\psi_2(\mathbf{r_2}) \Big] \, d^3\mathbf{r}_1 d^3\mathbf{r}_2 \end{aligned} \tag{4.8}$$

Since $\mathbf{r}_1$ and $\mathbf{r}_2$ are simply variables of integration in Eq. 4.8, the first and second terms

are equal, as are the third and fourth terms. Thus the matrix element can be written as

$$M_{if} = M_d - M_e \tag{4.9}$$

where

$$M_d = \int \psi^*_{1'}(\mathbf{r_1})\psi^*_{2'}(\mathbf{r_2})V\psi_1(\mathbf{r_1})\psi_2(\mathbf{r_2})d^3\mathbf{r}_1 d^3\mathbf{r}_2 \tag{4.10}$$

$$M_e = \int \psi^*_{2'}(\mathbf{r_1})\psi^*_{1'}(\mathbf{r_2})V\psi_1(\mathbf{r_1})\psi_2(\mathbf{r_2})d^3\mathbf{r}_1 d^3\mathbf{r}_2 \tag{4.11}$$

The terms M_d and M_e are referred to as the *direct* and *exchange* matrix elements respectively (which is labelled direct and which exchange is arbitrary). The impact ionisation perturbation operator, V, is the screened Coulomb potential [66,102]

$$V(\mathbf{r}_1, \mathbf{r}_2) = \frac{e^2}{4\pi\epsilon_0\epsilon(\mathbf{q},\omega)|\mathbf{r}_1 - \mathbf{r}_2|} \tag{4.12}$$

where the dielectric function, $\epsilon(\mathbf{q},\omega)$, is in general a complex function of the energy transfer, $\hbar\omega = E_1 - E_{1'}$, and wavevector transfer, $\mathbf{q} = \mathbf{k}_1 - \mathbf{k}_{1'}$, that occurs during the scattering event — see §4.2.3. To evaluate the matrix elements, the electron wavefunctions may be expressed as expansions in terms of plane waves:

$$\psi_\alpha = \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}_\alpha} A_\alpha(\mathbf{G}_\alpha) e^{i(\mathbf{k}_\alpha + \mathbf{G}_\alpha)\cdot\mathbf{r}} \tag{4.13}$$

where $\alpha = 1$, 2, 1′ or 2′ and Ω is the crystal volume. Then the direct matrix element in Eq. (4.10) becomes

$$M_d = \sum_{\mathbf{G}_1,\mathbf{G}_2,\mathbf{G}_{1'},\mathbf{G}_{2'}} \frac{e^2}{4\pi\epsilon_0\epsilon(\mathbf{q},\omega)\Omega^2} A^*_{1'}(\mathbf{G_{1'}})A^*_{2'}(\mathbf{G_{2'}})A_1(\mathbf{G_1})A_2(\mathbf{G_2}) \times \int e^{i\left[(\mathbf{G}_1 - \mathbf{G}_{1'} + \mathbf{k}_1 - \mathbf{k}_{1'})\cdot\mathbf{r}_1 + (\mathbf{G}_2 - \mathbf{G}_{2'} + \mathbf{k}_2 - \mathbf{k}_{2'})\cdot\mathbf{r}_2\right]} \frac{d^3\mathbf{r}_1 d^3\mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2|}. \tag{4.14}$$

The integral of Eq. (4.14) can be evaluated using the result that [103]

$$\int_\Omega \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} e^{i[\mathbf{K}_1\cdot\mathbf{r}_1 + \mathbf{K}_2\cdot\mathbf{r}_2]} d^3\mathbf{r}_1 d^3\mathbf{r}_2 = \frac{4\pi\Omega}{|\mathbf{K}_1|^2}\delta_{\mathbf{K}_1+\mathbf{K}_2,0} \tag{4.15}$$

Thus, using Eq. (4.15) in Eq. (4.14) we get

$$M_d = \sum_{\mathbf{G}_1,\mathbf{G}_2,\mathbf{G}_{1'},\mathbf{G}_{2'}} \frac{e^2}{\epsilon_0 \epsilon(\mathbf{q},\omega)\Omega} A^*_{1'}(\mathbf{G_{1'}}) A^*_{2'}(\mathbf{G_{2'}}) A_1(\mathbf{G_1}) A_2(\mathbf{G_2}) \times \frac{\delta_{\mathbf{G}_1+\mathbf{G}_2-\mathbf{G}_{1'}-\mathbf{G}_{2'}+\mathbf{k}_1+\mathbf{k}_2-\mathbf{k}_{1'}-\mathbf{k}_{2'},0}}{|\mathbf{G}_1 - \mathbf{G}_{1'} + \mathbf{k}_1 - \mathbf{k}_{1'}|^2} \quad (4.16)$$

where $\mathbf{q} = \mathbf{G}_1 - \mathbf{G}_{1'} + \mathbf{k}_1 - \mathbf{k}_{1'}$, $\hbar\omega = E_1 - E_{1'}$ and the Kronecker delta function leads to the conservation of crystal momentum to within a reciprocal lattice vector. A similar expression can be obtained for the exchange matrix element, M_e, but for brevity, only the direct expression will be considered in what follows. Appendix B discusses the form of the exchange matrix element in more detail.

4.2.1 Commonly Neglected Terms

The terms in Eq. (4.16) can be divided into two types, T_1 and T_2. The terms of type T_1 are those for which $\mathbf{G}_1 = \mathbf{G}_{1'}$ and so the denominator of the T_1-terms is $|\mathbf{k}_1 - \mathbf{k}_{1'}|^2$. The remaining terms T_2 are those for which $\mathbf{G}_1 \neq \mathbf{G}_2$ and therefore the denominator of these terms is $|\mathbf{G} + \mathbf{k}_1 - \mathbf{k}_{1'}|$. Because the T_1-terms have the smaller denominators, they tend to contribute more to the sum in Eq. (4.16) than the T_2-terms.

As will be discussed in §4.3, the impacting electron must have an energy above the conduction band edge at least equal to the band gap in order to excite an electron from the valence band. It follows that for wide band gap semiconductors, the electron of minimum energy to required cause impact ionisation will generally be higher in the conduction band than for narrow band gap semiconductors. This in turn means that in the wide band gap case, the impacting electron will normally have a larger $\mathbf{k}$-vector than in the narrow band gap case. For example, $In_{0.53}Ga_{0.47}As$ has a band gap of about half that of GaAs, and consequently impact ionisation in $In_{0.53}Ga_{0.47}As$ can be initiated from $\mathbf{k}$-states lying much closer to the Γ-point than in GaAs, as can be seen from Figs. 4.8 and 4.9 (which are discussed fully in §4.3).

Thus, in a narrow band gap semiconductor, the value of $\mathbf{k}_1 - \mathbf{k}_{1'}$ is small — much

smaller than the smallest reciprocal lattice vector. The result is that the terms of type T_1 dominate because their denominator is particularly small when $\mathbf{G}_1 - \mathbf{G}_{1'} = 0$. In this case, it is possible to neglect the terms of type T_2, reducing the expression in Eq. (4.16) to [102]

$$M_d \simeq \frac{e^2}{\epsilon_0 \epsilon(\mathbf{q},\omega) q^2 \Omega} \sum_{\mathbf{G}_1} A^*_{1'}(\mathbf{G}_1) A_1(\mathbf{G}_1) \sum_{\mathbf{G}_2} A^*_{2'}(\mathbf{G}_2) A_2(\mathbf{G}_2) \tag{4.17}$$

which is simply proportional to the product of the overlaps between the Bloch periodic parts of each particle's initial and final wavefunction,

$$M_d \simeq \frac{e^2}{\epsilon_0 \epsilon(\mathbf{q},\omega) \Omega} (\psi_{1'}|\psi_1)\,(\psi_{2'}|\psi_2). \tag{4.18}$$

where $(\psi_\alpha|\psi_\beta)$ denotes the overlap of the Bloch periodic parts of the wavefunctions ψ_α and ψ_β. Thus for narrow band gap semiconductors, the terms T_2 can be neglected due to their negligible contribution to the matrix element, and will be referred to here as the *Commonly Neglected Terms* or CNTs [82].

In a wide band gap semiconductor, the magnitude of $\mathbf{k}_1 - \mathbf{k}_{1'}$ is generally a significant fraction of the magnitude of the smallest reciprocal lattice vectors, and so the contribution to the matrix element of the terms in T_2 is important. In this case, the expression in Eq. (4.18) is not a good approximation and the full summation, as given by Eq. (4.16) must be used. Brand and Abram [82] have calculated that for the threshold impact ionisation transition of the form shown in Fig. 4.1 for GaAs, the correction to $|M_{if}|^2$ obtained by including the CNTs is about 50%.

Unfortunately, the use of the full sum given in Eq. (4.16), instead of the approximation given in Eq. (4.18) to calculate the matrix element, involves a considerable increase in computational effort. The sum in Eq. (4.16) has $O(N^2)$ times as many terms as Eq. (4.18), where N is the number of plane waves used in the basis set to expand the wavefunctions (65 in this work). In §4.2.4 it will be shown that by factorising Eq. (4.16), this can be reduced by a factor of N . Nevertheless, for wide band gap semiconductors, evaluation of the matrix element is more computationally intensive

than for narrow gap semiconductors where the approximate matrix element can be used without significant error.

4.2.2 Umklapp Terms

The terms of Eq. (4.16) can be divided into 'normal' and 'umklapp' terms. This division is separate to the division discussed above into types T_1 and T_2 — terms in T_1 can be both normal or umklapp, as can terms in T_2. The Kronecker delta function of Eq. (4.16) ensures that

$$\mathbf{k}_{1'} + \mathbf{k}_{2'} = \mathbf{k}_1 + \mathbf{k}_2 + \mathbf{G} \tag{4.19}$$

i.e. that crystal momentum is conserved to within a reciprocal lattice vector. With all the $\mathbf{k}$-vectors chosen so as to lie in the first Brillouin zone, normal processes are those for which $\mathbf{G} = 0$ and thus conserve crystal momentum exactly. Umklapp processes are the remaining terms for which $\mathbf{G} \neq 0$ and therefore conserve crystal momentum only to within a reciprocal lattice vector.

Some authors (e.g. [59,104]) use the term 'umklapp' to denote the terms of type T_2 described in the previous section. In this work, 'umklapp' will be used only to describe transitions for which $\mathbf{G} \neq 0$ in Eq. (4.19), and the terms in T_2 will be referred to as the CNTs.

4.2.3 The Dielectric Function

The dielectric function of the crystal, ϵ, appears in the impact ionisation perturbation operator and hence in the transition matrix element for the process. Generally, ϵ is a complex number which is a function of wave vector and frequency: $\epsilon = \epsilon(\mathbf{q}, \omega)$. In this case, $\mathbf{q}$ is interpreted as the momentum transfer in the process and ω is the energy transfer [66].

In wide band gap semiconductors, where both energy and momentum transfer are

likely to be significant, accurate calculations of impact ionisation rates require the use of the full **q**- and ω-dependent expression for the dielectric function.

4.2.4 Factorisation of Matrix Element Summation

In the evaluation of the expression for M_d in Eq. (4.16) one of the four summations can be carried out immediately by virtue of the Kronecker delta which ensures the conservation of crystal momentum to within a reciprocal lattice vector, i.e.

$$\mathbf{G}_1 + \mathbf{G}_2 - \mathbf{G}_{1'} - \mathbf{G}_{2'} + \mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_{1'} - \mathbf{k}_{2'} = 0 \tag{4.20}$$

Summing over $\mathbf{G}_2$ gives

$$M_d = \frac{e^2}{\epsilon_0 \Omega} S_d \tag{4.21}$$

where S_d contains the terms that are functions of the summation indices:

$$S_d = \sum_{\mathbf{G}_1, \mathbf{G}_{1'}, \mathbf{G}_{2'}} \frac{A^*_{1'}(\mathbf{G_{1'}}) A^*_{2'}(\mathbf{G_{2'}}) A_1(\mathbf{G_1}) A_2(\mathbf{G_2})}{\epsilon(\mathbf{q}, \omega) q^2} \tag{4.22}$$

and

$$\mathbf{G}_2 = \mathbf{G}_{1'} + \mathbf{G}_{2'} - \mathbf{G}_1 + \mathbf{G}_u \tag{4.23}$$

where

$$\mathbf{G}_u = \mathbf{k}_{1'} + \mathbf{k}_{2'} - \mathbf{k}_1 - \mathbf{k}_2 \tag{4.24}$$

This sum, S, can be factorised as follows [104,105]. A vector $\mathbf{G}_\Delta$ is defined as

$$\mathbf{G}_\Delta = \mathbf{G}_1 - \mathbf{G}_{1'} \tag{4.25}$$

which gives $\mathbf{G}_2 = \mathbf{G}_{2'} - \mathbf{G}_\Delta + \mathbf{G}_u$ and $\mathbf{G}_{1'} = \mathbf{G}_1 - \mathbf{G}_\Delta$. S is then summed over the

indices $\mathbf{G}_1$, $\mathbf{G}_{2'}$ and $\mathbf{G}_\Delta$.

$$S_d = \sum_{\mathbf{G}_1,\mathbf{G}_{2'},\mathbf{G}_\Delta} \frac{A^*_{1'}(\mathbf{G}_1 - \mathbf{G}_\Delta)A^*_{2'}(\mathbf{G}_{2'})A_1(\mathbf{G}_1)A_2(\mathbf{G}_{2'} - \mathbf{G}_\Delta + \mathbf{G}_u)}{\epsilon(\mathbf{q},\omega)q^2} \tag{4.26}$$

which can be factorised as

$$S_d = \sum_{\mathbf{G}_\Delta} \left\{ \Big[\sum_{\mathbf{G}_1} A^*_{1'}(\mathbf{G}_1 - \mathbf{G}_\Delta)A_1(\mathbf{G}_1) \Big] \Big[\sum_{\mathbf{G}_{2'}} A^*_{2'}(\mathbf{G}_{2'})A_2(\mathbf{G}_{2'} - \mathbf{G}_\Delta + \mathbf{G}_u) \Big] \frac{1}{\epsilon(\mathbf{q},\omega)q^2} \right\}. \tag{4.27}$$

Note that Eqs. (4.22) and (4.26) are exactly equivalent expressions only if we use an infinite number of plane waves in the basis set (i.e. there is an infinite number of **G**-vectors in each sum). In the finite basis set used, the change of indices discards some terms. However, only the terms corresponding to the largest values of q are lost, and the error is negligible. The sum over three indices in Eq. (4.26) has $O(N^3)$ terms in it (N being the number of plane waves in the basis set — 65 in this case), while the sum in Eq. (4.27) has only $O(N^2)$ terms, making it much quicker to evaluate. This is an important consideration since the matrix element will be required at a large number of points in **k**-space during a typical rate calculation. Eq. (4.27) is therefore used to obtain $|M_{if}|$.

As with Eq. (4.16), we can identify the two types of terms T_1 and T_2 in Eq. (4.27). Terms for which $\mathbf{G}_\Delta = 0$ correspond to T_1, and in this case Eq. (4.27) leads directly to Eq. (4.17). Terms for which $\mathbf{G}_\Delta \neq 0$ correspond to T_2, the CNTs. Normal and umklapp terms correspond to $\mathbf{G}_u = 0$ and $\mathbf{G}_u \neq 0$ respectively.

4.2.5 Mixed Spin States

As already discussed in §2.3.2, the wavefunctions are in general not pure spin-up or spin-down, but a linear combination of the two. The wavefunction can be expressed as

$$\psi(\mathbf{r}) = {}^{\uparrow}\psi(\mathbf{r})|\uparrow\rangle + {}^{\downarrow}\psi(\mathbf{r})|\downarrow\rangle. \tag{4.28}$$

where $|\uparrow\rangle$ and $|\downarrow\rangle$ represent orthonormal spin-up and spin-down eigenstates respectively. Putting Eq. (4.28) into Eq. (4.10) we get

$$\begin{aligned} M_d = \int \Big[& {}^{\uparrow}\psi^*_{1'}(\mathbf{r}_1){}^{\uparrow}\psi^*_{2'}(\mathbf{r}_2)V{}^{\uparrow}\psi^*_{1}(\mathbf{r}_1){}^{\uparrow}\psi^*_{2}(\mathbf{r}_2) \\ & + {}^{\uparrow}\psi^*_{1'}(\mathbf{r}_1){}^{\downarrow}\psi^*_{2'}(\mathbf{r}_2)V{}^{\uparrow}\psi^*_{1}(\mathbf{r}_1){}^{\downarrow}\psi^*_{2}(\mathbf{r}_2) \\ & + {}^{\downarrow}\psi^*_{1'}(\mathbf{r}_1){}^{\uparrow}\psi^*_{2'}(\mathbf{r}_2)V{}^{\downarrow}\psi^*_{1}(\mathbf{r}_1){}^{\uparrow}\psi^*_{2}(\mathbf{r}_2) \\ & + {}^{\downarrow}\psi^*_{1'}(\mathbf{r}_1){}^{\downarrow}\psi^*_{2'}(\mathbf{r}_2)V{}^{\downarrow}\psi^*_{1}(\mathbf{r}_1){}^{\downarrow}\psi^*_{2}(\mathbf{r}_2) \Big] d^3\mathbf{r}_1 d^3\mathbf{r}_2 \end{aligned} \tag{4.29}$$

The exchange matrix element contributes another four terms. Numerically, each of these terms is integrated separately, using Eq. (4.27) with the appropriate spin part of the wavefunction ${}^{\uparrow}\psi_\alpha(\mathbf{r})$ or ${}^{\downarrow}\psi_\alpha(\mathbf{r})$ substituted in place of $\psi_\alpha(\mathbf{r})$, and the amplitudes summed to obtain the complete matrix element.

A more detailed discussion of the form of the matrix element summation, including the spin terms for the direct and exchange matrix elements, is given in Appendix B.

4.2.6 Convergence of M_{if} with respect to N

The wavefunctions are represented as in Eq. (4.13) using a basis set of N plane waves. A total of $2N$ terms are needed to represent both the spin-up and spin-down parts of the wavefunction. Fig 4.6 shows how impact ionisation matrix elements calculated for InGaAs and SiGe converge as a function of N. The ordinate shows the average squared magnitude of matrix elements calculated for a large number transitions which are chosen randomly, but are all energy and momentum conserving. In the case of InGaAs, the matrix element is slightly sensitive to the number of plane waves even when as many as 307 are used. However, the average value obtained with 65 plane waves (as used in this work) is within about 2% of the value obtained with 307. For SiGe, in which the matrix elements are generally significantly lower than in InGaAs, the convergence is much poorer, with 307 plane waves not being sufficient to ensure good convergence. However, the number of plane waves that can be used is limited

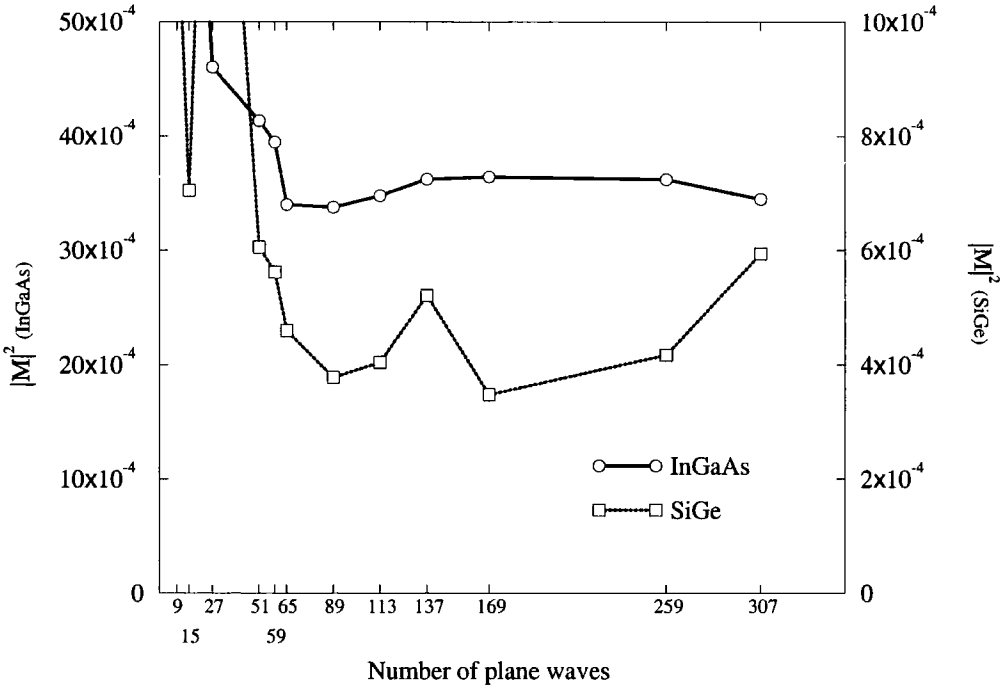


Figure 4.6: The convergence of impact ionisation matrix elements in InGaAs and SiGe as a function of N — the number of plane waves in the expansion of the wavefunctions. The matrix elements for each material are in the same (arbitrary) units. Note the different scales used for the ordinate in each case.

by the available computer processing power, and so in this work 65 are used, as for InGaAs. This gives matrix elements that are in error with respect to those obtained using 307 plane waves by about 30%.

4.3 Impact Ionisation Thresholds

During the process of impact ionisation, the impacted electron is given sufficient energy to be excited from the valence band to the conduction band. In an electron initiated process, this energy is supplied by the impacting electron which undergoes a transition in the conduction band from a state of high energy to one of low energy. It is therefore a minimum requirement that the impacting electron be at least E_g above the conduction band edge, where E_g is the band gap energy. In fact, the additional requirement of

momentum conservation means that in general the impacting electron must have a significantly greater energy.

The range of **k**-states from which an electron can cause impact ionisation is therefore limited. Anderson and Crowell have developed a method to determine the location in **k**-space of the thresholds [106]. However it has been shown to obtain incorrect thresholds under certain conditions [20]. Therefore a method due to Beattie [107,108] is used here. We begin by defining an *energy difference function.*

4.3.1 The Energy Difference Function

Although energy is conserved in an allowed impact ionisation process[c], it is still possible to consider formally transitions in which there is a change of energy and define an energy difference function.

For any single impact ionisation process, from states $\mathbf{k}_1$ and $\mathbf{k}_2$ to states $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$, the energy difference function is defined as

$$\Delta E(\mathbf{k}_1, \mathbf{k}_{1'}, \mathbf{k}_{2'}) = \Big[E(\mathbf{k}_{1'}) + E(\mathbf{k}_{2'})\Big] - \Big[E(\mathbf{k}_1) + E(\mathbf{k}_{1'} + \mathbf{k}_{2'} - \mathbf{k}_1 + \mathbf{G})\Big] \qquad (4.30)$$

where the impacted electron state $\mathbf{k}_2$ is expressed in terms of the other vectors so as to ensure crystal momentum is conserved to within a reciprocal lattice vector. Energy conservation is satisfied when

$$\Delta E(\mathbf{k}_1, \mathbf{k}_{1'}, \mathbf{k}_{2'}) = 0 \qquad (4.31)$$

Ignoring energy conservation for the moment, we can treat ΔE simply as a function of three **k**-vectors. For any value of $\mathbf{k}_1$ (the state of the impacting electron), there is a combination of the two remaining vectors (the final states) for which ΔE is a minimum. Therefore this minimum value ΔE_{min} is a function of $\mathbf{k}_1$ only. Similarly the maximum value of the energy difference function, ΔE_{max}, is also a function of $\mathbf{k}_1$ only.

[c] As discussed in the introduction to this chapter, energy can be transferred to or from the electron system if other agencies such as phonons are also involved but such processes are not considered here.

Since the energy in each band is a smooth function of wavevector, ΔE is also a smoothly varying function of the final state vectors $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$. Therefore if $\Delta E_{min} < 0$ and $\Delta E_{max} > 0$ there must be some combination of $\mathbf{k}_1$ and $\mathbf{k}_2$ for which $\Delta E = 0$, corresponding to an energy conserving transition. Momentum is conserved automatically in the Eq. (4.30) by expressing $\mathbf{k}_2$ in terms of the other three vectors and so we can say that, if for a given state $\mathbf{k}_1$ in a given band,

$$\Delta E_{min} < 0 < \Delta E_{max} \tag{4.32}$$

then impact ionisation can be initiated by a carrier in that state.

When $\mathbf{k}_1$ and $\mathbf{k}_{1'}$ are in the same band, ΔE_{max} is always positive. When $\mathbf{k}_{1'}$ is in a lower band than $\mathbf{k}_1$, this is also usually the case, and it is sufficient in most cases only to test that $\Delta E_{min} < 0$ to determine whether impact ionisation can be initiated from a given state.

4.3.2 Thresholds and Anti-thresholds

For $\mathbf{k}_1$ near the bottom of the conduction band, the value of ΔE_{min} will be positive, since impact ionisation cannot be caused by low energy electrons. As $\mathbf{k}_1$ moves away from the band edge, ΔE_{min} changes as a function of $\mathbf{k}_1$. Fig. 4.7 illustrates an example variation of $\Delta E_{min}(\mathbf{k}_1)$.

The point on the $\mathbf{k}_1$-axis at which ΔE_{min} changes from positive to negative values (marked **T** in Fig. 4.7) is known as the *threshold* for the process. It is at this point that impact ionisation becomes possible. Further along the $\mathbf{k}_1$-axis in the example shown, the value of ΔE_{min} passes back through zero to become positive. This point (marked **A** on the diagram) is known as the *anti-threshold.* Here, due to the nature of the band structure and the difficulty of simultaneously satisfying energy and momentum conservation, impact ionisation becomes impossible once again.

Thresholds and anti-thresholds for the first conduction bands of GaAs and InGaAs are shown in Figs. 4.8 and 4.9 respectively. In each figure, the base of the plot is the

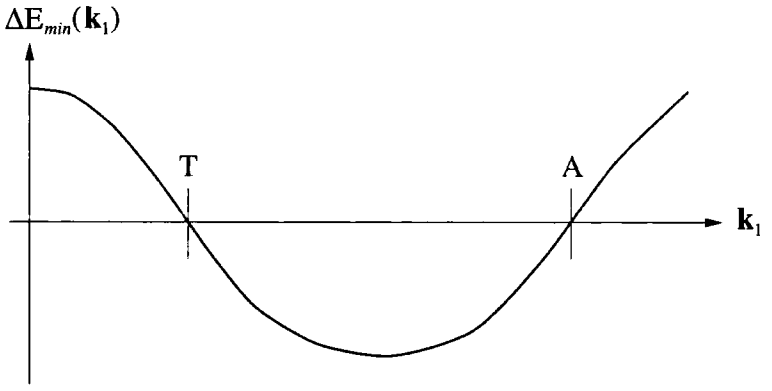


Figure 4.7: The minimum value of ΔE, defined in Eq. (4.30), as a function of the impacting vector, $\mathbf{k}_1$. Where $\Delta E_{min} < 0$, impact ionisation is possible. The point marked **T** is the *Threshold*, and the point marked **A** is the *Anti-threshold.*

$k_z = 0$ plane of the Brillouin zone, while the vertical axis represents the energy of the conduction band as a function of (k_x, k_y). The bands are coloured green where impact ionisation can be caused by an electron at that wavevector and red otherwise.

The band gap of GaAs is about 1.5 eV, and so electrons must gain at least this energy (and in practice more) to cause impact ionisation. The thresholds for GaAs, shown in Fig. 4.8, are therefore located high in the first conduction band. In this part of the Brillouin zone, the $E(\mathbf{k})$ relation for the band is highly anisotropic, and this is reflected in the anisotropy of the thresholds.

InGaAs has a much smaller band gap than GaAs of about 0.75 eV, and the thresholds in this material, shown in Fig. 4.9, are at correspondingly lower energy in the first conduction band. The threshold (the transition from red to green on moving out from the Γ-minimum) is more isotropic, being located in a more spherically-symmetric part of the band structure. Anti-thresholds can also be seen in the [100] and [110] directions as a transition from green to red as the edge of the Brillouin zone is approached.

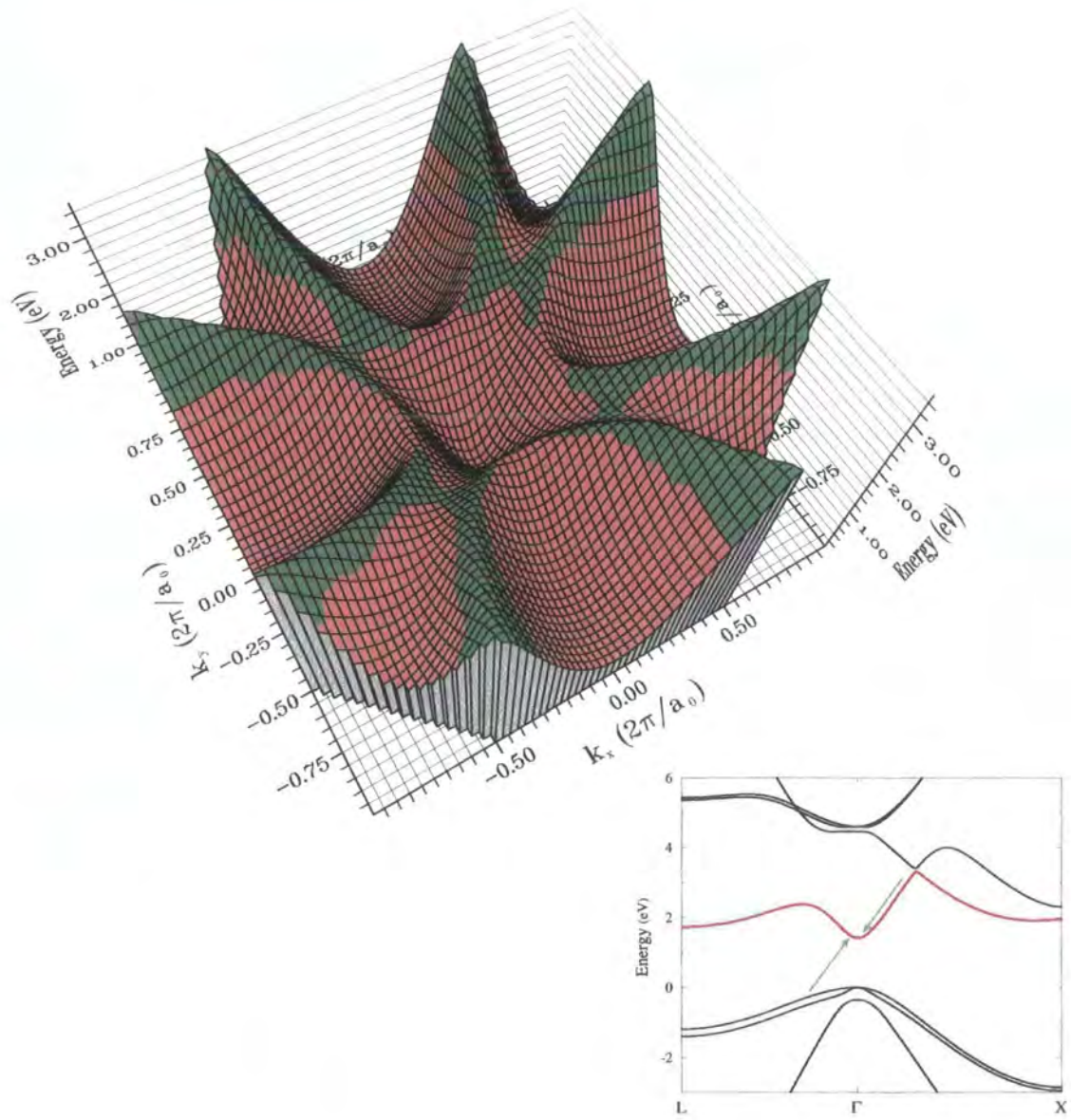


Figure 4.8: Thresholds in GaAs (at $T = 300K$). The base of the upper plot is the $k_z = 0$ plane of the Brillouin zone and the height is the energy of the 1st conduction band. The plot is coloured green at **k**-points from which impact ionisation can be initiated and red where impact ionisation is impossible. The lower plot shows the process for which the threshold is calculated.

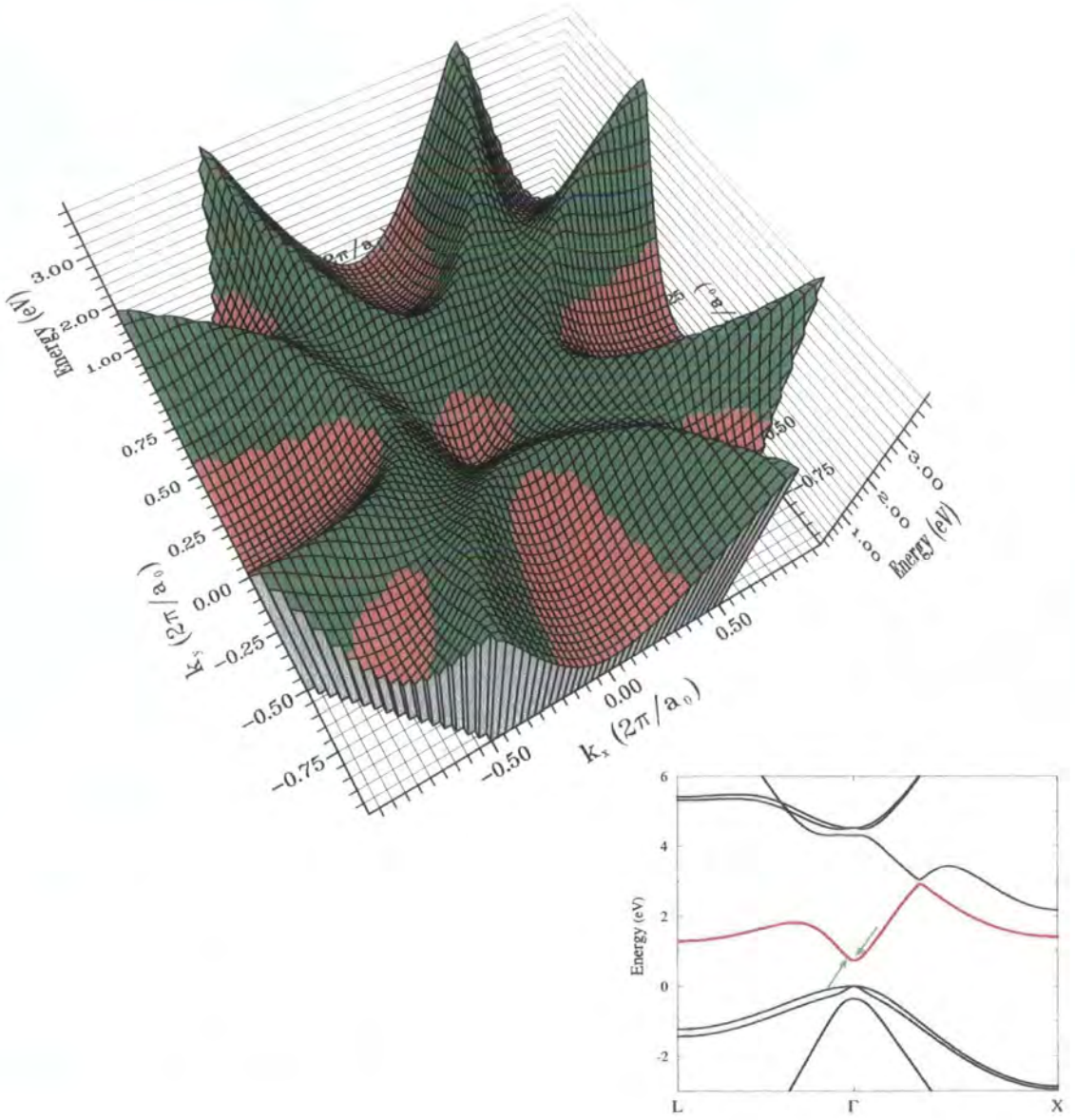


Figure 4.9: Thresholds in $In_{0.53}Ga_{0.47}As$ (at $T = 0K$). The plot is of the same form as that in Fig. 4.8. The band gap of $In_{0.53}Ga_{0.47}As$ is approximately half that of GaAs, hence the much greater range of states for which impact ionisation is possible. Note also that the threshold is closer to the Γ-point, the significance of which is discussed in §4.2.1.

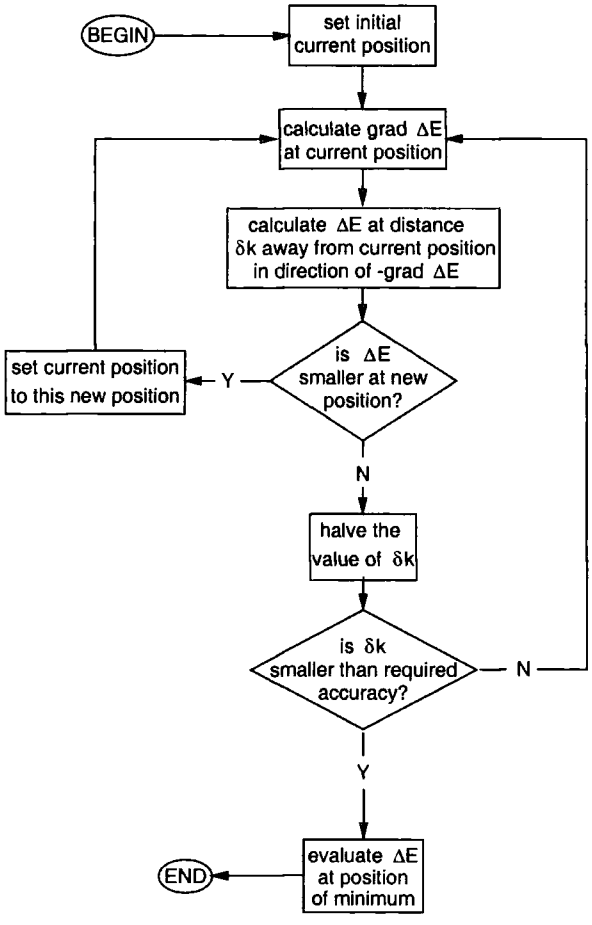


Figure 4.10: The algorithm to determine if a given impacting carrier can initiate impact ionisation, by finding the minimum of the energy difference function, ΔE_{min}. If it is less than zero, impact ionisation can occur from the given impacting state. To find thresholds, the Brillouin zone must be sampled throughout its volume and ΔE_{min} evaluated for each initiating **k**-point.

4.3.3 Finding Thresholds

To determine whether a state can initiate impact ionisation, we must search for the position of the minimum of the energy difference function. This is performed by an algorithm which 'walks' down the gradient of the ΔE function in 6-dimensional $\mathbf{k}_{1'}$,$\mathbf{k}_{2'}$-space in decreasing step lengths as the minimum is approached. This algorithm is represented in Fig. 4.10.

There is a difficulty in that the function $\Delta E(\mathbf{k}_{1'}, \mathbf{k}_{2'})$ will generally have several local minima for a given impacting vector, and the algorithm will find only one of them from

a given starting point in $\mathbf{k}_{1'}$,$\mathbf{k}_{2'}$-space. The local minimum thus obtained may not be the absolute minimum of the function as required, and so the search algorithm must be tried from several starting points to ensure that the absolute minimum is found.

In the case of electron initiated transitions with final states in the first conduction band, the search is initialised with $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$ located at the bottom of the Γ-, X- and L-valleys — all combinations of valley pairs are taken. For hole initiated transitions with the impacted hole states in the first conduction band, one final state is initialised to the top of the valence band (i.e. the bottom of the band in terms of hole energies) and the other final state initialised either to the top of the valence band also, or at such a position that the impacted carrier is located at the Γ-, X- or L-valley bottom.

4.3.4 The Condition of Equal Velocities

It has been shown that $\Delta E_{min}(\mathbf{k}_1) = 0$ at thresholds and anti-thresholds. It can also be shown that the group velocities associated with the states $\mathbf{k}_2$, $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$ are equal there [106]. The proof starts by demonstrating that the result is true for any minimum of the energy difference function, whatever its value.

Hence we first seek to prove that if for a given state of the impacting electron $\mathbf{k}_1$, the final states $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$ are varied so that the energy difference function $\Delta E(\mathbf{k}_1, \mathbf{k}_{1'}, \mathbf{k}_{2'})$ is a minimum, the group velocities at each of the states $\mathbf{k}_2$, $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$ will be equal.

Consider small variations made in $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$. To conserve crystal momentum, we must have

$$0 + d\mathbf{k}_2 = d\mathbf{k}_{1'} + d\mathbf{k}_{2'} \tag{4.33}$$

where the zero on the left hand side is due to the fact that the impacting state is fixed, so $d\mathbf{k}_1 = 0$. The change in the $\mathbf{k}$-vectors, affects the energies of those states. The change in ΔE is given by

$$d(\Delta E) = d\mathbf{k}_{1'} \cdot \nabla_k E_{1'} + d\mathbf{k}_{2'} \cdot \nabla_k E_{2'} - d\mathbf{k}_2 \cdot \nabla_k E_2 \tag{4.34}$$

but since we are considering the case when ΔE is minimised, small changes in $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$ lead to no change in ΔE. Hence,

$$0 = d\mathbf{k}_{1'} \cdot \nabla_k E_{1'} + d\mathbf{k}_{2'} \cdot \nabla_k E_{2'} - d\mathbf{k}_2 \cdot \nabla_k E_2. \tag{4.35}$$

Using the definition of group velocity $\mathbf{v}$

$$\mathbf{v} = \frac{1}{\hbar} \nabla_k E \tag{4.36}$$

we can write Eq. (4.35) as

$$0 = d\mathbf{k}_{1'} \cdot \mathbf{v}_{1'} + d\mathbf{k}_{2'} \cdot \mathbf{v}_{2'} - d\mathbf{k}_2 \cdot \mathbf{v}_2 \tag{4.37}$$

which, using Eq. (4.33), can be re-written as

$$0 = d\mathbf{k}_{1'} \cdot (\mathbf{v}_{1'} - \mathbf{v}_2) + d\mathbf{k}_{2'} \cdot (\mathbf{v}_{2'} - \mathbf{v}_2). \tag{4.38}$$

Since $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$ can be varied independently, each of the terms in parentheses in Eq. (4.38) must be zero, and so we have

$$\mathbf{v}_{1'} = \mathbf{v}_{2'} = \mathbf{v}_2. \tag{4.39}$$

That is, the group velocities of the state of the impacted electron and the final states are equal at the point at which ΔE is minimised. In particular the result holds for the cases where $\Delta E_{min} = 0$ which are the thresholds and anti-thresholds of the impact ionisation process.

4.4 The Rate Integration

The rate given by Fermi's Golden Rule in Eq. (4.3) is the transition rate for one single transition from initial states $\mathbf{k}_1$ and $\mathbf{k}_2$ to final states $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$. To obtain the *total* rate of transitions for an impacting electron at $\mathbf{k}_1$, it is necessary to sum the rates for all possible individual transitions corresponding to distinct processes. Thus, the total

rate is given by

$$R_{II}(\mathbf{k}_1) = \sum_{\mathbf{k}_{1'},\mathbf{k}_{2'}} \frac{2\pi}{\hbar} |M_{if}|^2 \, \delta(E_{1'} + E_{2'} - E_1 - E_2). \tag{4.40}$$

where it is assumed that all states in the conduction band are unoccupied and all states in the valence band are occupied. Note that the sum is over the final states only because the remaining state $\mathbf{k}_2 = \mathbf{k}_{1'} + \mathbf{k}_{2'} - \mathbf{k}_1 + \mathbf{G}$ is determined by the conservation of crystal momentum.

It is convenient to convert the sum over discrete $\mathbf{k}$-states to an integral

$$R_{II}(\mathbf{k}_1) = \frac{\Omega^2}{(2\pi)^6} \int \frac{2\pi}{\hbar} |M_{if}|^2 \, \delta(E_{1'} + E_{2'} - E_1 - E_2) \, d^6\mathbf{k} \tag{4.41}$$

which is performed over the six-dimensional volume containing all pairs of final states $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$. Note that the volume element is written here as $d^6\mathbf{k}$ rather than $d^3\mathbf{k}_{1'} d^3\mathbf{k}_{2'}$ to emphasise the fact that the integral is performed over six independently variable coordinates. For the purposes of the integration, no significance need be attached to the fact that these six coordinates are in fact two sets of coordinates in three-dimensional space.

The matrix element M_{if} in Eq. (4.41) is replaced with the expression in Eq. (B.11)[d]. Writing $E_{1'} + E_{2'} - E_1 - E_2$ as ΔE, we get

$$R_{II}(\mathbf{k}_1) = \frac{e^4}{32\pi^5 \epsilon_0^2 \hbar} \int |S|^2 \, \delta(\Delta E) \, d^6\mathbf{k} \tag{4.42}$$

Care must be taken in the interpretation of the expression 'all pairs of final states', over which the integral is performed and this will be discussed in §5.5 of Chapter 5.

[d] Which is the complete version of the expression of Eq. (4.21) for the direct matrix element.

Chapter 5

Impact Ionisation: Numerical Integration

The expression for the impact ionisation rate given in Chapter 4,

$$R_{II}(\mathbf{k}_1) = \frac{e^4}{32\pi^5\epsilon_0^2\hbar}\int |S|^2\,\delta(\Delta E)\,d^6\mathbf{k} \tag{5.1}$$

is an integral over a 6-dimensional volume in $\mathbf{k}$-space, but the Dirac delta function ensures that only points on the surface satisfying the condition $\Delta E(\mathbf{k}_1, \mathbf{k}_{1'}, \mathbf{k}_{2'}) = 0$ contribute to the result. Numerically, the integral can be treated in two ways: as a volume or a surface integral.

To treat it as a volume integral, we must relax the requirement imposed by the delta function that energy be conserved exactly. The delta function is replaced with a top-hat function of finite width [58] which ensures that energy is conserved to within some suitably small value. Instead of lying on a surface, the final states corresponding to allowed transitions now lie within a shell of finite volume. The integral in Eq. (5.1) can then be performed by evaluating the integrand at a large number of randomly chosen pairs of final states. Only final states lying within the volume defined by the top-hat function contribute to the result. The width of this top-hat must be small enough to ensure that the original delta function is accurately approximated, but large

enough that a statistically significant number of sampled points lie within the volume it defines.

The alternative approach is to convert the volume integral of Eq. (5.1) into an integral over the surface for which $\Delta E = 0$ is satisfied exactly [61]. This eliminates the problem of having to choose a suitable width for the top-hat function. However, the difficulty now lies in determining the position in $\mathbf{k}_{1'}$,$\mathbf{k}_{2'}$-space of the $\Delta E = 0$ surface, which may be complicated.

In this work, both volume and surface methods are used to evaluate rates. The two methods have different advantages and disadvantages, as will be discussed in §5.5.1, and so the use of both provides a convenient check on the accuracy of the numerical calculations. The implementation of these methods is discussed in subsequent sections.

5.1 Numerical Volume Integration

Replacing the Dirac delta function of Eq. (5.1) with a top-hat function of width $2\delta e$, we obtain the following expression for the rate:

$$R_{II}(\mathbf{k}_1) = \frac{e^4}{32\pi^5\epsilon_0^2\hbar}\int_{\Omega_0} I_v \, d^6\mathbf{k} = \frac{e^4}{32\pi^5\epsilon_0^2\hbar}\,\overline{I_v}\,\Omega_{BZ}^2 \tag{5.2}$$

where

$$I_v(\mathbf{k}_{1'},\mathbf{k}_{2'}) = \begin{cases} \frac{1}{2\delta e}|S|^2 & \text{if } |\Delta E| \le \delta e, \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

and $\overline{I_v}$ is the mean value of I_v throughout the hyper-volume Ω_0 over which the integration is to be done. This volume, which will be referred to in what follows as the 'joint-Brillouin zone',contains all pairs of points $(\mathbf{k}_{1'},\mathbf{k}_{2'})$, such that $\mathbf{k}_{1'}$ and $\mathbf{k}_{2'}$ lie in the familiar Brillouin zone of 3-dimensional $\mathbf{k}$-space[a]. The problem then is to find the value of $\overline{I_v}$ which is achieved in this work using a Monte Carlo algorithm.

[a]Care must be taken to include each pair of states only once in the integration — see §5.5

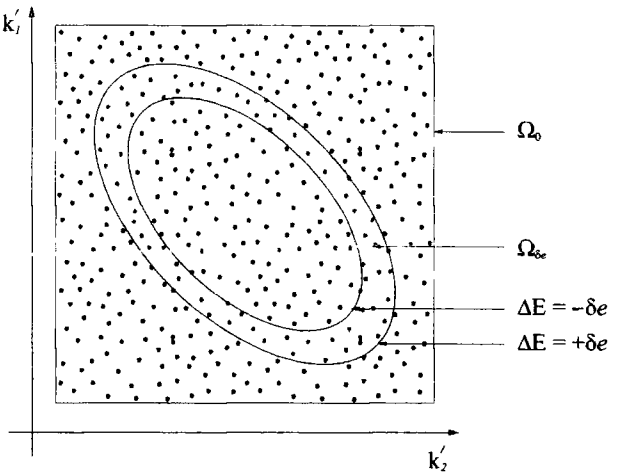


Figure 5.1: A 2-dimensional representation of a simple 6-dimensional volume integration algorithm. The square is the volume Ω_0 over which the integration is performed. The ellipses are the $\Delta E = \pm\delta e$ surfaces which enclose the volume $\Omega_{\delta e}$. The dots are random sampling points: the integrand is zero at the black dots and non-zero at the red dots. The statistical error on the integral is a function only of the number of red sampling points.

5.1.1 A Simple Integration Algorithm

A simple approach to obtaining $\overline{I_v}$ would be to pick coordinates randomly throughout the volume Ω_0 and at each evaluate I_v given by Eq. (5.3) — and hence calculate the mean [109]. Fig. 5.1 illustrates the essential features of the procedure. The square represents the volume to be sampled, Ω_0, and the ellipses represent the $\Delta E = \pm\delta e$ surfaces which enclose volume contributing to the integral, $\Omega_{\delta e}$. The integrand at sampling points lying outside $\Omega_{\delta e}$ (marked black) is zero and these do not contribute to the rate. The red points lying inside $\Omega_{\delta e}$ correspond to transitions which conserve energy to within $\pm\delta e$, and these contribute to the total rate. A sufficiently large number of points must be picked so that the statistical error on the value of $\overline{I_v}$ is reduced to some required tolerance.

Two factors affect the numerical accuracy of the algorithm. Firstly, the value of δe must be sufficiently small that the Dirac delta function of Eq. (5.1) is well approximated. Secondly, the statistical noise on the final value of $\overline{I_v}$ is reduced only by points

for which the integrand is non-zero (i.e. the red points of Fig. 5.1). We can pick a very large number of points throughout Ω_0, but if only a few lie inside the volume $\Omega_{\delta e}$, the statistical error on $\overline{I_v}$ will be large.

These two considerations work against one another. If we reduce the value of δe, the run-time of the rate integration program is increased due to the need to pick more sampling points to obtain a given statistical error on the result. Conversely, increasing δe reduces the run-time necessary, but increases the error due to the poor approximation of the energy conserving delta function. It turns out that using the computers available for this work, the value of δe could not be chosen to approximate the delta function sufficiently well without increasing the computational requirements beyond a practical level. Thus the simple rate integration algorithm described above cannot be used. In the next section, a more sophisticated algorithm is described which reduces this problem.

5.1.2 A Better Integration Algorithm

As described above, the problem to be overcome is the fact that the volume for which the integrand is non-zero is much smaller than the volume to be sampled, i.e. $\Omega_{\delta e} \ll \Omega_0$. Thus very few randomly sampled points contribute to the total rate and the statistical error on the result is high. One approach to solving this problem is to restrict the sampling points to some volume Ω_B which is much smaller than Ω_0 but which nevertheless completely encloses $\Omega_{\delta e}$. The Monte Carlo integration is performed in the same way as for the simple algorithm described in §5.1.1. However, because $\Omega_{\delta e}$ now constitutes a much greater fraction of Ω_B than of Ω_0, far more sampled points correspond to positions of non-zero integrand, and the convergence of the result with respect to the number of points sampled is correspondingly more rapid.

As before, the total rate is calculated from the mean value of the integrand. The quantity returned by the above algorithm is $\overline{I'_v}$ — the mean value of the integrand within Ω_B. To get the rate, we require the mean value of the integrand throughout Ω_0,

which is given by

$$\overline{I_v} = \overline{I'_v} \times \frac{\Omega_B}{\Omega_0} \tag{5.4}$$

from which the total rate is obtained using Eq. (5.2).

The method of reduction of the sampled volume from Ω_0 to Ω_B is described in the next section.

5.1.3 Reduction of the Volume to be Sampled

The reduction of the volume to be sampled from Ω_0 to Ω_B is performed by identifying parts of Ω_0 that do not contain any of the volume $\Omega_{\delta e}$. It is an iterative procedure which begins by dividing Ω_0 into several 'sub-volumes'. The sub-volumes not containing any of the region $\Omega_{\delta e}$ are discarded. Those remaining go on to the next iteration in which they are themselves each divided into sub-volumes. The process of dividing and discarding is repeated B times, the volume remaining at the end being Ω_B. The algorithm is represented schematically in Fig. 5.2.

Diagram A of Fig. 5.2 represents the initial state which is the same as for the simple algorithm in Fig. 5.1. Ω_0 is represented by the square, and $\Omega_{\delta e}$ by the region lying between the ellipses.

Diagram B represents the situation after the first iteration. The initial volume Ω_0 has been bisected in each direction to form a set of sub-volumes. In the 2-dimensional representation of the diagram, four sub-volumes are formed; in 6-dimensions, bisection in every direction results in 64 sub-volumes. In the diagram, all the sub-volumes contain part of $\Omega_{\delta e}$ and so all are kept.

Diagram C corresponds to the state after two bisections. The unshaded sub-volumes do not contain any part of $\Omega_{\delta e}$ and have been discarded. The shaded sub-volumes are retained for the next iteration.

Diagram D shows the shaded sub-volumes remaining after the fourth iteration. The volume they occupy is considerably smaller than the original Ω_0 but nevertheless

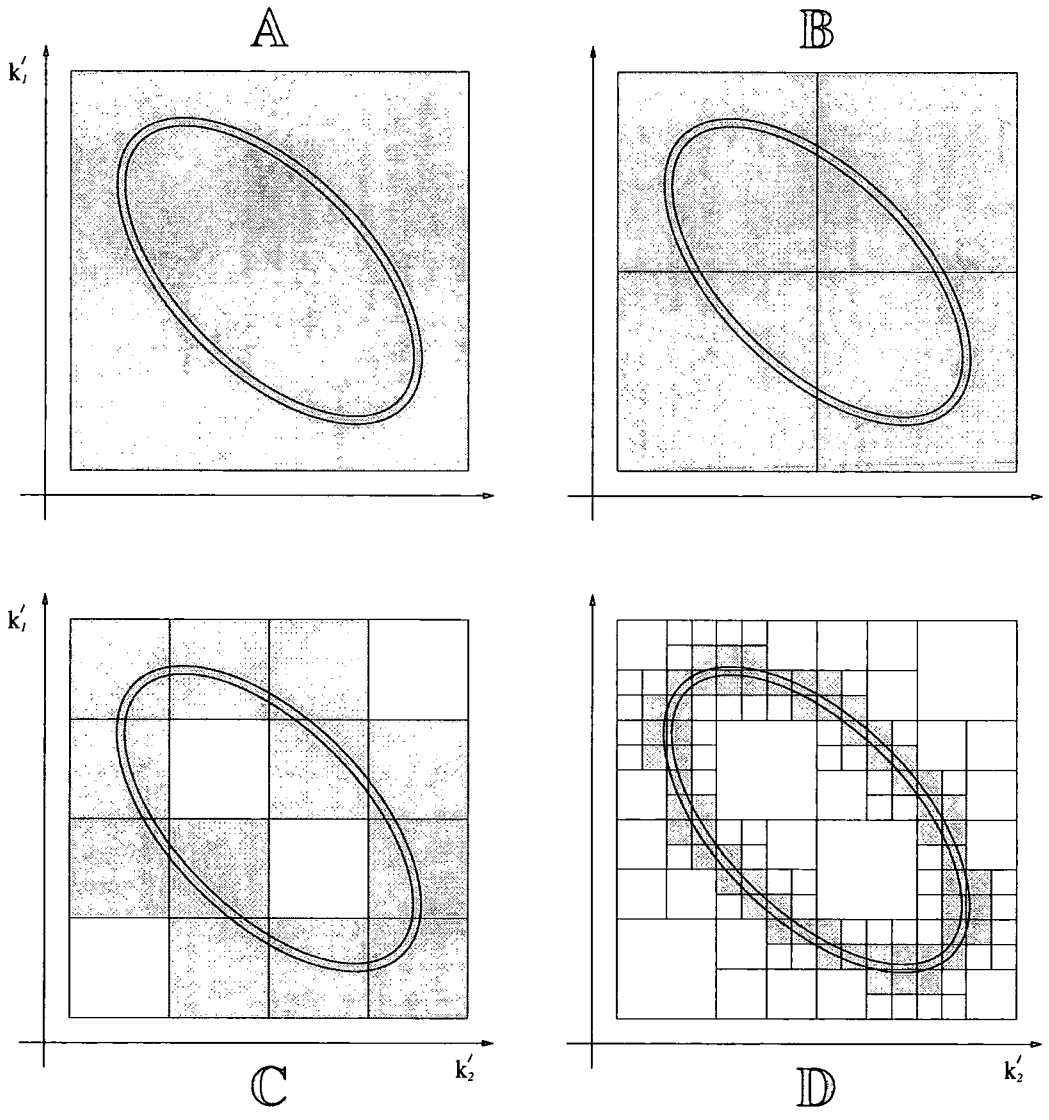


Figure 5.2: A 2-dimensional representation of the better 6-dimensional volume integration algorithm (compare with the simple algorithm of Fig. 5.1). The square in Diag. A is the volume Ω_0 over which the integration is performed and $\Omega_{\delta e}$ is the volume of interest, between the ellipses. In Diags. B–D, Ω_0 is iteratively divided into sub-volumes, and those not containing $\Omega_{\delta e}$ discarded.

completely contains $\Omega_{\delta e}$ between the ellipses marking the $\Delta E = \pm\delta e$ surfaces.

A summary of the notation that will be used in describing the volume reduction algorithm is now given.

- The iterations are numbered with the index b. The initial state, shown by Diagram A in Fig. 5.2, corresponds to the 0^{th} iteration, and the final iteration is the B^{th}, i.e. $0 \leq b \leq B$. The b^{th} iteration is complete after b bisections and discards are complete.

- The volume remaining after b iterations is labelled Ω_b. Hence, Ω_0 is the initial integration volume, Ω_B the volume remaining after the reduction phase is complete, and: $\Omega_0 \geq \Omega_1 \geq \ldots \geq \Omega_b \geq \ldots \geq \Omega_B > \Omega_{\delta e}$.

- The number of sub-volumes remaining after b iterations is labelled N_b. If no sub-volumes are discarded then $N_{b+1} = 64N_b$. In all but the earliest iterations, sub-volumes can be discarded and $N_{b+1} < 64N_b$.

- The initial volume Ω_0 is taken to be a hyper-cube centred at the origin of $\mathbf{k}_1$,$\mathbf{k}_2$-space and of side length 2 (in units of $\frac{2\pi}{a}$). The actual integration volume — the joint-Brillouin zone — is smaller than, and contained within this cube. When the Monte Carlo integration is performed, the integrand at points in the hyper-cube lying outside the double-Brillouin zone is taken to be zero. Since Ω_0 is cubic, all sub-volumes are cubic. The side length of sub-volumes formed after b iterations is 2^{1-b} (in units of $\frac{2\pi}{a}$)

The actual number of bisection steps used in the algorithm can be adjusted to give acceptable accuracy of the integration — B should be chosen as large as possible to minimise interpolation errors, but memory requirements increase rapidly with increasing B (see §5.1.4 and §5.1.6). After only a few bisections ($\lesssim 10$) the volume Ω_B which must now be sampled will in many cases be smaller than the original volume Ω_0 by a factor exceeding 10^7.

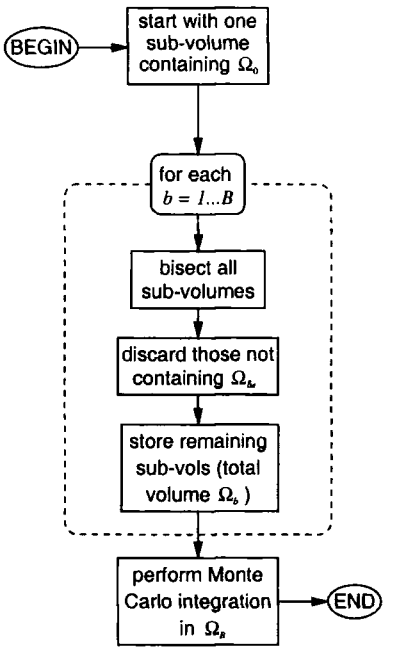


Figure 5.3: A graphical representation the better volume integration algorithm.

It should be noted that the major time saving achieved by this algorithm is as a result of discarding sub-volumes at each bisection iteration. The number of sub-volumes that would be obtained from eight bisections without discarding any is $N_8 = 64^8$, i.e. more than 10^{14}. Determining in which of these the volume $\Omega_{\delta e}$ lay would be no more efficient than performing the integral using the simple algorithm of §5.1.1. However, by using the above algorithm, the tiny fraction of sub-volumes containing part of the $\Omega_{\delta e}$ volume can be located without having to consider the vast majority directly. Fig. 5.3 summarises this bisection algorithm.

5.1.4 Discarding Sub-Volumes

To implement the algorithm described in the previous section, a method of quickly determining which sub-volumes can be discarded is required. Given a particular sub-volume, one possible method is as follows.

Within the sub-volume, the maximum and minimum values of the energy difference function $\Delta E(\mathbf{k}_{1'}, \mathbf{k}_{2'})$ are determined[b]. As was noted in §4.3.1, if $\Delta E_{min} \leq 0 \leq \Delta E_{max}$,

[b]These maximum and minimum energies should not be confused with stationary points — most sub-volumes will not contain a stationary point.

then the sub-volume must contain the surface $\Delta E = 0$. Similarly, if $\Delta E_{min} \leq +\delta e$ and $\Delta E_{max} \geq -\delta e$ then it must contain the volume $|\Delta E| \leq \delta e$, i.e. the volume $\Omega_{\delta e}$. Thus we have the rule that a sub-volume must be kept if

$$\Delta E_{min} \leq +\delta e \qquad \text{and} \qquad \Delta E_{max} \geq -\delta e \tag{5.5}$$

and discarded otherwise.

Searching 6-dimensional phase space for the positions of ΔE_{min} and ΔE_{max} in every sub-volume is impractically time consuming, and so the condition given above for keeping or discarding sub-volumes is not applied directly.

Instead the following energies are defined:

$$E^{min} = E^{min}_{1'} + E^{min}_{2'} - E_1 - E^{max}_2 \tag{5.6}$$

$$E^{max} = E^{max}_{1'} + E^{max}_{2'} - E_1 - E^{min}_2 \tag{5.7}$$

where $E^{min}_{1'} \ldots E^{max}_{1'}$ is the range of energies for states $\mathbf{k}_{1'}$ within the sub-volume, $E^{min}_{2'} \ldots E^{max}_{2'}$ the range of energies for $\mathbf{k}_{2'}$, and $E^{min}_2 \ldots E^{max}_2$ the range of energies of the corresponding impacted states, $\mathbf{k}_2$. It follows that within any sub-volume it must always be the case that

$$E^{min} \leq \Delta E_{min} < \Delta E_{max} \leq E^{max}. \tag{5.8}$$

Thus, we can adjust the rule for keeping or discarding sub-volumes to the following: a sub-volume is kept if

$$E^{min} \leq +\delta e \qquad \text{and} \qquad E^{max} \geq -\delta e \tag{5.9}$$

and discarded otherwise. Because of Eq. (5.8), the above rule will never discard a sub-volume the would have been kept using Eq. (5.5). Therefore the use of Eq. (5.9) in place of Eq. (5.5) will not affect the result of the integration.

The new rule is not as efficient at reducing the volume to be sampled, as it will keep some sub-volumes that could have been discarded using the old rule. However, as

the side length of the sub-volumes becomes smaller with each bisection step, E^{min} and E^{max} tend towards ΔE_{min} and ΔE_{max} respectively, and the conditions in Eqs. (5.5) and (5.9) converge.

The advantage of using the new rule is that we must now search three independent 3-dimensional functions (i.e. $E_{1'}(\mathbf{k}_{1'})$, $E_{2'}(\mathbf{k}_{2'})$ and $E_2(\mathbf{k}_2)$) for their maxima and minima instead of the 6-dimensional function $\Delta E(\mathbf{k}_{1'}, \mathbf{k}_{2'})$. Searching a 3-dimensional function for maxima and minima is not significantly easier than searching a 6-dimensional one, and in that sense little has been gained by adopting the new rule for keeping and discarding sub-volumes. However the advantage of the 3-dimensional rule is that all the necessary maxima and minima can be pre-calculated and stored. It only remains to retrieve their values during the rate integration, which can be done very rapidly.

Pre-calculation of Energy Maxima and Minima

The 3-dimensional Brillouin zone is divided into a grid, labelled G_B, of equal cubes each having a side length of 2^{1-B} (in units of $\frac{2\pi}{a}$), i.e. the same side length as the 6-dimensional sub-volumes created as a result of the final bisection iteration. Energy is assumed to vary linearly within each cube and hence the maximum energy in a cube is taken to be the maximum energy of its corners, and similarly for the minimum. Maximum and minimum energies in each cube are stored for each of the final and impacted state bands.

Fig. 5.4 represents schematically how the 3-dimensional mesh of cubes is related to the 6-dimensional sub-volumes of integration. A given 6-dimensional sub-volume created by the B^{th} iteration will contain final states ($\mathbf{k}_{1'}$,$\mathbf{k}_{2'}$). One of the 3-dimensional cubes of grid G_B will contain all the states $\mathbf{k}_{1'}$ in the sub-volume while another cube (or possibly the same one) will contain states $\mathbf{k}_{2'}$. By requiring that the impacting carrier wavevector lies at one of the nodes of G_B, the impacted carrier states $\mathbf{k}_2$ associated with the sub-volume in question will also be contained in eight of the cubes in G_B (eight because the impacted carrier states lie in a cubic volume of twice the side length

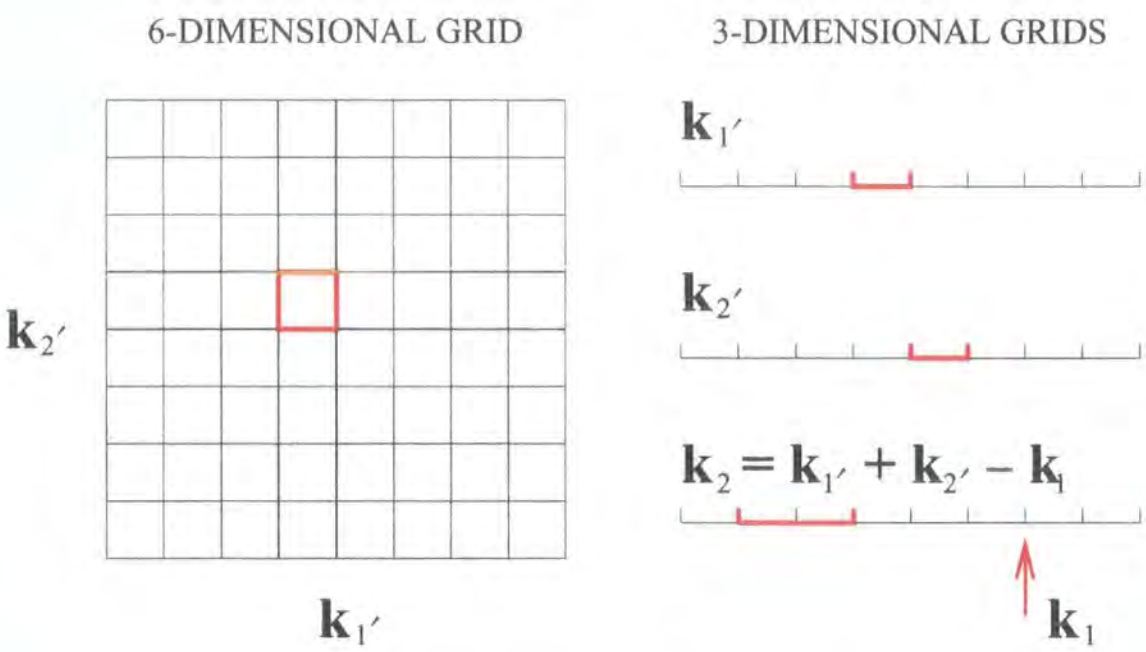


Figure 5.4: A 2-dimensional representation of the 6-dimensional grid used to discretise the final state phase space, and 1-dimensional representations of the corresponding 3-dimensional grids on which energy data is stored. A 6-dimensional element of final state space is shown in red. The energy data associated with this 6-d volume is stored in the corresponding 3-dimensional elements, also shown in red, of the impacted and final state grids.

of the 6-D sub-volume). Since the maximum and minimum energies of the bands of interest in each cube are stored, we can apply the rule of Eq. (5.9) simply by retrieving their values from memory, which is very rapid.

Similar grids G_b are constructed to correspond the the sub-volumes formed by each stage of bisection b, up to the 0^{th} grid which consists of a single cube surrounding the Brillouin zone. These coarser grids can be rapidly obtained: the maximum energy in a cube in grid G_b is straightforwardly obtained from the maxima of the eight cubes in grid G_{b+1} from which it is formed.

Several points are worth noting about the above procedure:

- The use of the 3-dimensional grids is essential. Storing maximum and minimum values of ΔE for 6-dimensional grids would allow application of Eq. (5.5) instead of the less efficient Eq. (5.9), but we would again have the problem of finding max-

ima and minima in the order of 10^{14} mesh cubes. The number of 3-dimensional grid cubes required is about seven orders of magnitude smaller.

- The requirement that impacting carrier wavevectors lie at the nodes of the finest grid G_B is generally not very restrictive. In a typical rate calculation B will be set to 7 or 8, which makes the spacing of G_B's nodes $\frac{1}{64}$ or $\frac{1}{128}$ (in units of $\frac{2\pi}{a}$) respectively.

- The memory required by G_b is $M_b \propto 8^b$. Thus the total memory M_{tot} required by all the grids $G_0 \ldots G_B$, increases very rapidly with B; $M_{tot} \propto 8^B$. Furthermore, the majority of the memory is used only by the finest grid; $M_B \simeq \frac{7}{8} M_{tot}$. However, because impacting carrier wavevectors are required to lie at the nodes of G_B, the symmetry of the Brillouin zone can used to advantage, and G_B is defined only within the irreducible wedge (discussed in §3.2). This reduces M_B by a factor of 48 and M_{tot} by a factor of about 8. Because the impacting carrier wavevector does not in general lie at the nodes of the coarser grids, $G_0 \ldots G_{B-1}$ must be defined throughout the Brillouin zone.

- A saving in run-time is achieved through the fact that the calculation of maximum and minimum energies in G_B need only be performed once for all impacting vectors (provided the combination of bands involved remains the same). Only the coarser grids need to be re-constructed for each impacting vector.

5.1.5 Storage of Sub-Volumes

During the volume integration, the positions throughout Ω_0 of all undiscarded sub-volumes must be stored. After a certain number of bisection iterations, the storage requirements will usually exceed the available memory (depending on how much of the phase space is occupied by $\Omega_{\delta e}$). Therefore an upper limit must be placed on the number of sub-volumes that will be stored during the integration.

Suppose that at the b^{th} iteration, the number of sub-volumes N_b retained by the rule of Eq. (5.9) exceeds the limit imposed, N_{max}. In this case, N_{max} of the N_b sub-volumes are chosen at random and stored for the next iteration, while the rest are discarded despite their being selected by Eq. (5.9). Thus, the fraction of the total volume that should be sampled that has actually been retained at this iteration is F_b, where

$$0 < \quad F_b = \frac{N_{max}}{N_b} \quad \leq 1 \tag{5.10}$$

It is assumed that provided N_{max} is large, the fraction of sub-volumes retained is a representative sample of all the sub-volumes that should have been kept. The volume Ω_B remaining at the end of B bisection stages, as a fraction of that which would remain if N_{max} were infinite, is F, given by

$$F = \prod_{b=1}^{B-1} F_b \tag{5.11}$$

The product is over all bisections except the B^{th} due to the fact that the sub-volumes formed at the last iteration do not need to be stored. As these final sub-volumes are created by bisection of their predecessors, the Monte Carlo integration is carried out in each and then they are discarded. Thus, F_B effectively has the value 1.

As in §5.1.2, the Monte Carlo integration returns the mean value of the integrand $\overline{I'_v}$ within the volume Ω_B. Because the fraction F of sub-volumes sampled is representative of all the sub-volumes that would be sampled for infinite N_{max}, $\overline{I'_v}$ is the same for this reduced set as it would be for the whole set. To obtain the mean value of the integrand for the whole integration volume $\overline{I_v}$, Eq. (5.4) must be replaced with the expression

$$\overline{I_v} = \overline{I'_v} \times \frac{\Omega_B}{\Omega_0} \times \frac{1}{F} \tag{5.12}$$

Finally the total rate $R_{II}(\mathbf{k}_1)$ is obtained from $\overline{I_v}$ using Eq. (5.2), as before.

5.1.6 Performance of the Volume Algorithm

The volume integration algorithm described above requires the setting of several adjustable parameters which are not directly connected with the physical aspects of the problem but determine the performance of the numerical algorithm. Therefore, we attempt to choose values for the parameters such that the final result is insensitive to small variations of the values. The parameters are summarised below, together with considerations to be made when choosing their values.

Top-hat function width, δe: The top-hat function is used to approximate the Dirac delta function that ensures conservation of energy, and therefore the smaller the value of δe, the better the approximation. Too small a value will cause large statistical errors in the final rate due to few sampled points lying in the volume in which it is non-zero.

Number of bisection stages, B: Since band energy is interpolated linearly in the sub-volumes created by the final bisection step, the sub-volume side length should be as small as possible to reduce interpolation errors. This means as many bisections as possible should be performed. However since memory use increases as 8^B, the number of bisections is limited by the available computational resources. We hope to be able to choose a sufficiently large value for B that the calculated rate converges. A higher value for B also allows more freedom in where impacting carrier wavevectors can be located.

Maximum number of sub-volumes stored, N_{max}: As with the number of bisection steps, the highest possible value of N_{max} should be chosen that is compatible with the available memory resources. Again, we hope to be able to choose a value sufficiently large that the rate has converged with respect to it.

Number of sampling points taken, N_{samp}: The more points sampled within the volume remaining after the bisection and reduction phase, the lower will be the

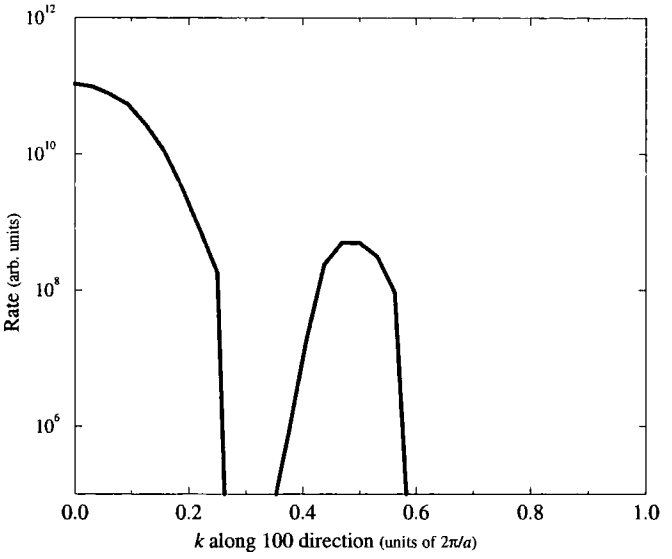


Figure 5.5: Calculated impact ionisation rate from the 2nd conduction band of GaAs plotted versus wavevector along the 100-direction using the values of the algorithm parameters given in Table 5.1.

statistical error on the rate. However, more sampling points require more matrix element evaluations and hence more computer time. Thus a suitable balance between accuracy and run-time must be found.

Fig. 5.5 shows the impact ionisation rate in GaAs as a function of position of the impacting carrier wavevector. The rate is calculated for transitions in which the initial states are in bands 12 and 8, and the final states are in band 9 (see Chapter 2, Table 2.2 for the band labelling notation). The abscissa corresponds to the position of the impacting carrier wavevector along the 100 direction, and the ordinate (logarithmic scale) is the corresponding rate (in arbitrary units). The rate calculation was performed with the parameters B, N_{max}, N_{samp} and δe set to values given in the 'Near threshold' column of Table 5.1, i.e. values for which the result of the calculation was converged for all impacting carrier wavevectors.

Figs. 5.6, 5.7, 5.8 and 5.9 indicate how the convergence occurs with respect to each of the parameters B, N_{max}, N_{samp} and δe. In each case, the vertical axis corresponds to

Parameter	Value Required for Convergence	
	Away from threshold	Near threshold
B	≥ 6	≥ 7
N_{max}	$\gtrsim 10^4$	$\gtrsim 10^4$
N_{samp}	$\gtrsim 10^3$	$\gtrsim 10^3$
δe (eV)	10^{-6}–10^{-1}	10^{-6}–10^{-2}

Table 5.1: Parameter settings for the volume integration giving well converged rates.

the logarithm of the transition rate, and one of the horizontal axes to the magnitude of $\mathbf{k}_1$ along the 100 direction, i.e. the same information as presented in Fig. 5.5. The other horizontal axis then corresponds to the variation of the relevant adjustable parameter.

The plots show that each parameter can be given values for which convergent rate results are obtained. For each parameter the highest rates (i.e. away from threshold) converge quickest, with lower rates (i.e. near threshold) converging slower. The relevant quantitative details are given in the figure captions and Table 5.1. In the case of parameters B, N_{max} and N_{samp} convergence improves as the parameter value is increased. Computational requirements (memory and/or CPU) also increase with these parameters, and so the smallest values for which acceptable convergence is achieved should be used. In the case of the δe, there is a 'window' of values which give converged results, with non-convergent rates being obtained if the parameter is set too low or too high.

5.2 Conversion of Integral from Volume to Surface

An alternative algorithm is one due to Beattie [61] which treats the rate calculation as a surface integral. Thus, where the integral described in §5.1 is performed throughout the *volume* defined by the approximate energy conservation condition $\Delta E \leq \delta e$, the integral here is performed over the *surface* defined by the exact energy conservation condition $\Delta E = 0$.

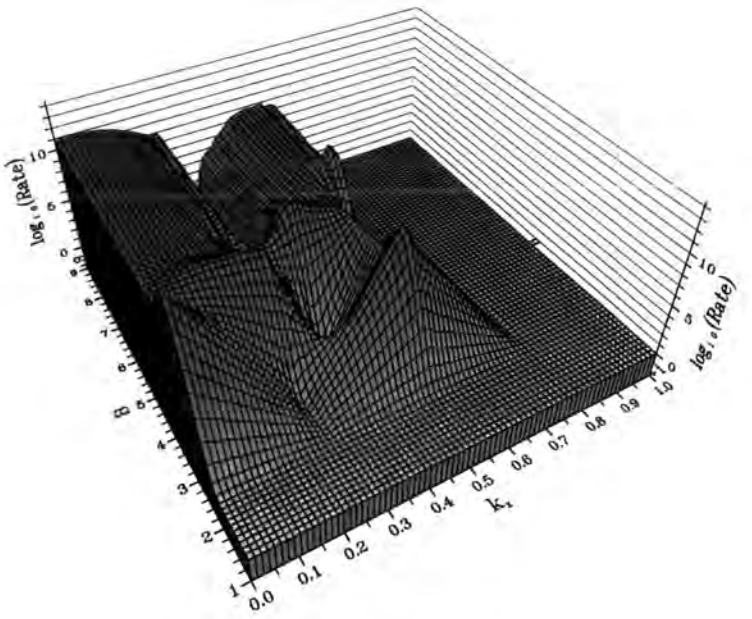


Figure 5.6: Convergence of the rate WRT B — the number of division iterations. The rate converges to the result of Fig. 5.5 for $B \gtrsim 7$. Note that number of impacting vectors at which the rate is calculated falls with decreasing B. Hence the rate at, for example, $B = 3$ can only be calculated at 5 impacting vectors.

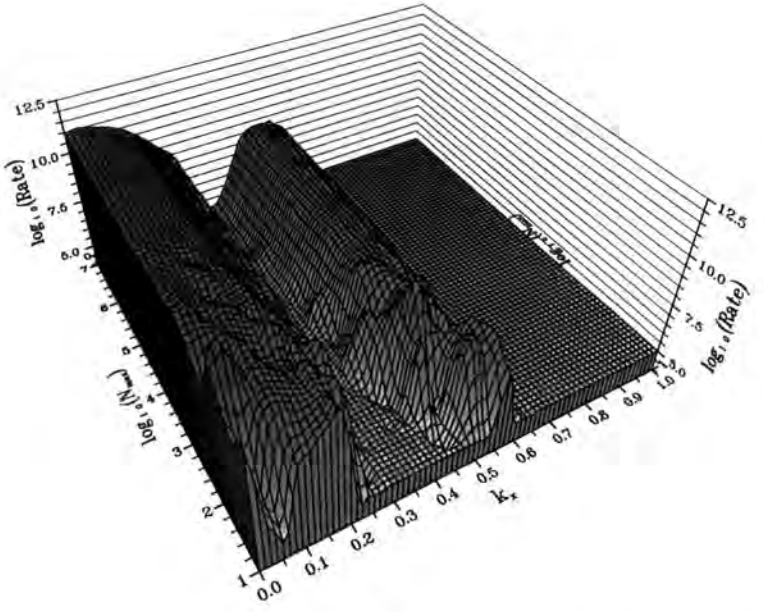


Figure 5.7: Convergence of the rate WRT N_{max} — the maximum number of sub-volumes stored. Statistical errors in the calculated rate decrease as N_{max} increases. Good convergence is achieved for $N_{max} \gtrsim 10^4$.

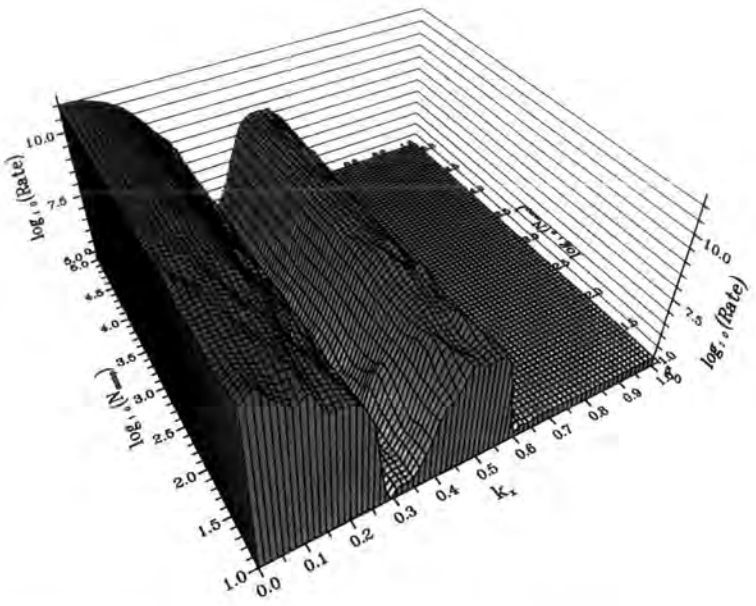


Figure 5.8: Convergence of the rate WRT N_{samp} — the number of random sampling points taken. As with N_{max}, the statistical error on the result decreases as N_{samp} increases, with good convergence being obtained for $N_{samp} \gtrsim 10^3$.

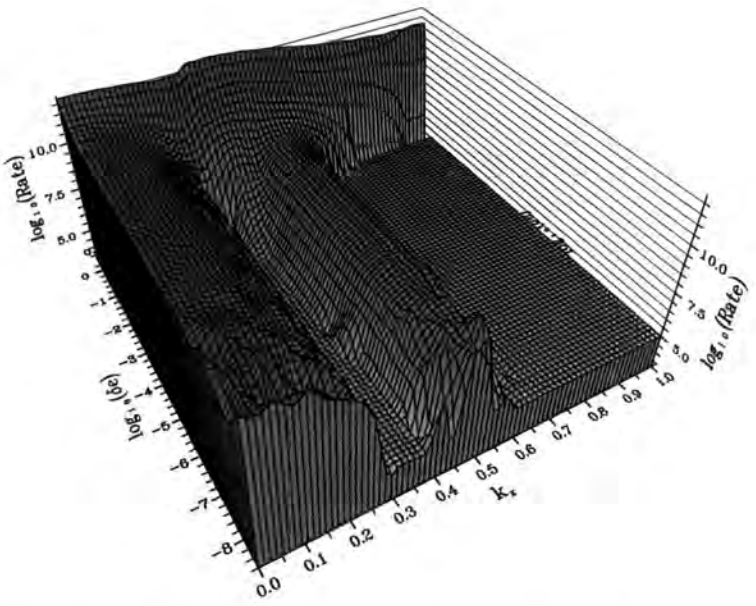


Figure 5.9: Convergence of the rate WRT δe — the width of the the top-hat function. Too large a value for δe poorly approximates the Dirac delta function, while too small a value leads to large statistical errors in the result (and also problems related to machine precision). Convergence is achieved for $10^{-6} \lesssim \delta e \lesssim 10^{-2}$ eV.

The volume integral in $\mathbf{k}_{1'}$,$\mathbf{k}_{2'}$-space of Eq. (5.1) is converted into an integral over the $\Delta E = 0$ surface in the following way. The volume element $d^6\mathbf{k}$ can be written in terms of polar coordinates as

$$d^6\mathbf{k} = k^5 dk d\Theta \tag{5.13}$$

where dk is an element of length in the radial direction and $d\Theta$ is an element of 6-dimensional solid angle. Thus, in polar coordinates, Eq. (5.1) is written as

$$R_{II}(\mathbf{k}_1) = \frac{e^4}{32\pi^5\epsilon_0^2\hbar} \int_\Theta \left\{ \int_k |S|^2 \, \delta(\Delta E) \, k^5 dk \right\} d\Theta. \tag{5.14}$$

The integral with respect to k in braces can be carried out first:

$$\int_k |S|^2 \, \delta(\Delta E) \, k^5 dk \quad = \quad |S|^2 \, k^5 \left| \frac{d(\Delta E)}{dk} \right|^{-1} \Bigg|_{\Delta E=0} \tag{5.15}$$

in which the Dirac delta function has picked out the value of the integrand at the radial coordinate where $\Delta E = 0$. Finally, putting Eq. (5.15) into Eq. (5.14) gives the surface integral

$$R_{II}(\mathbf{k}_1) = \frac{e^4}{32\pi^5\epsilon_0^2\hbar} \int_\Theta \left(|S|^2 \, k^5 \left| \frac{d(\Delta E)}{dk} \right|^{-1} \right) \Bigg|_{\Delta E=0} d\Theta \tag{5.16}$$

Note that in [61], Beattie describes the application of his algorithm to analytic band structure. Here, the application of his algorithm to pseudopotential band structure is similar to that described by Wilson *et al* [64].

5.3 Numerical Surface Integration

The surface integral is carried out using a numerical method analogous to that described in §5.1 for the volume integral. Writing the element of solid angle as [61]

$$d\Theta = \sin^3\theta_1 \, \sin^2\theta_2 \, \sin\theta_3 \, \mathrm{d}(\cos\theta_1) \, \mathrm{d}(\cos\theta_2) \, \mathrm{d}(\cos\theta_3) \, \mathrm{d}(\cos\theta_4) \, \mathrm{d}\theta_5 \tag{5.17}$$

the integral of Eq. (5.16) will be carried out with respect to the set of coordinates $(\cos\theta_1, \ldots, \cos\theta_4, \theta_5)$ throughout the 6-dimensional solid angle Θ_T, defined by [61]

$$\begin{aligned} -1 \leq \quad \cos\theta_i \quad &\leq 1 \qquad (i = 1 \ldots 4) \\ 0 \leq \quad \theta_5 \quad &\leq 2\pi \end{aligned} \tag{5.18}$$

Eq. (5.17) is substituted into Eq. (5.16) giving

$$\begin{aligned} R_{II}(\mathbf{k}_1) &= \frac{e^4}{32\pi^5\epsilon_0^2\hbar} \int_{\Theta_T} I_s \; \mathrm{d}(\cos\theta_1) \; \mathrm{d}(\cos\theta_2) \; \mathrm{d}(\cos\theta_3) \; \mathrm{d}(\cos\theta_4) \; \mathrm{d}\theta_5 \\ &= \frac{e^4}{32\pi^5\epsilon_0^2\hbar} \; \overline{I_s} \; \Theta_T \end{aligned} \tag{5.19}$$

where

$$I_s(\cos\theta_1, \ldots, \cos\theta_4, \theta_5) = \left(|S|^2 \; k^5 \left| \frac{d(\Delta E)}{dk} \right|^{-1} \right) \Bigg|_{\Delta E = 0} \sin^3\theta_1 \; \sin^2\theta_2 \; \sin\theta_3 \tag{5.20}$$

and $\overline{I_s}$ is the mean value of the integrand throughout the whole solid angle Θ_T over which the integral is to be done. As with the volume integral, the problem now is to calculate the mean value $\overline{I_s}$, which is done using a Monte Carlo algorithm.

5.3.1 The Integration Algorithm

The evaluation of $\overline{I_s}$ is carried out using a similar algorithm to that described in §5.1.1 to evaluate $\overline{I_v}$. In the volume algorithm, coordinates $(\mathbf{k}_{1'}, \mathbf{k}_{2'})$ were picked at random throughout the volume Ω_0 and hence $\overline{I_v}$ calculated. Similarly, here coordinates $(\cos\theta_1, \ldots, \cos\theta_4, \theta_5)$ are picked at random throughout Θ_T and hence $\overline{I_s}$ is calculated.

Fig. 5.10 schematically represents the algorithm. The thick line represents the $\Delta E = 0$ surface in $\mathbf{k}_{1'}$,$\mathbf{k}_{2'}$-space. The coordinate origin has been moved to the point marked **O** on the diagram, which is the position at which $\Delta E = \Delta E_{min}$ (see §4.3.3). This position always lies *inside* the surface.

A set of coordinates $(\cos\theta_1, \ldots, \cos\theta_4, \cos\theta_5)$ have been picked at random, the corresponding radial direction being marked on the diagram. The intersection of this

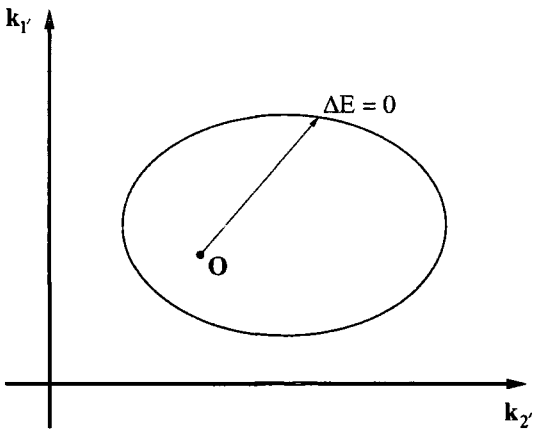


Figure 5.10: A schematic representation of the surface of allowed transitions in $\mathbf{k}_{1'}$,$\mathbf{k}_{2'}$-space. The coordinate origin is moved to **O** — the point at which ΔE is a minimum — which lies inside the surface. From this origin, the surface is sampled in randomly chosen radial directions to determine the average value of the integrand over it.

radial line with the $\Delta E = 0$ surface is determined, and the integrand I_s evaluated at this point. By repeating this evaluation for many randomly chosen sets of angular coordinates, the average value of the integrand, $\overline{I_s}$, is determined, to within some statistical error.

5.3.2 Inclusion of the Whole Surface

The surface shown in Fig. 5.10 is relatively easy to integrate over. However, in regions of the Brillouin zone where the band structure is complicated, the surface of allowed transitions is likely to be correspondingly complicated. Several surfaces may exist for a given initiating electron, and these surfaces may join. Fig. 5.11 shows how the $\Delta E = 0$ surface may become complicated as the impacting electron gains energy above threshold.

Figs. 5.11**a**, **b** and **c** show the surface(s) of allowed transitions for increasing impacting electron energy. Fig. 5.11**a** shows the case just above threshold: only one surface of allowed transitions exists, and it is of the form shown in Fig. 5.10.

In Fig. 5.11**b** the electron has gained sufficient energy as to be able to access two

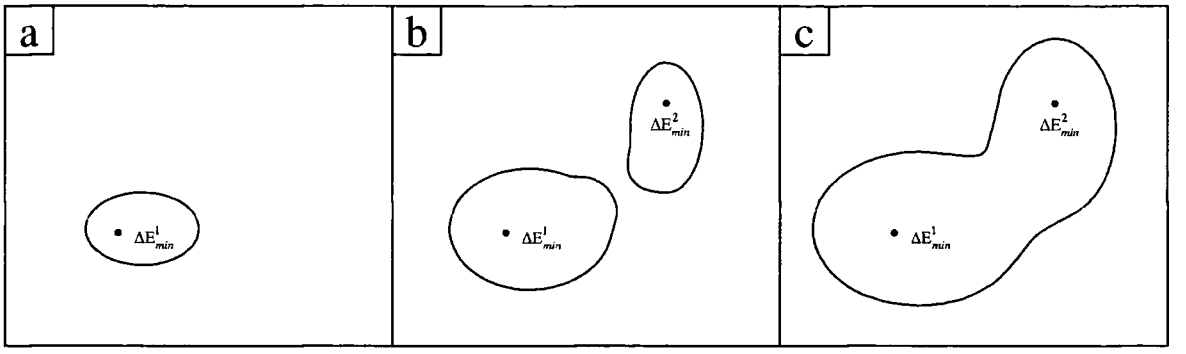


Figure 5.11: The surface of allowed transitions in $\mathbf{k}_{1'}$,$\mathbf{k}_{2'}$-space, growing as the impacting electron gains energy above threshold. In Fig. **a**, the initiating electron is just above threshold and the surface is simple. In Fig. **b**, at higher energy, two surfaces are accessible, each one contributing to the rate. In Fig. **c**, at still higher energy, the two surfaces have joined to form a single complicated surface that poses problems for integration.

surfaces of allowed transitions. Each is of the form of the simple surface in Fig. 5.11a and each is integrated over separately from origins ΔE^1_{min} and ΔE^2_{min}. The total rate for the impacting carrier is the sum of the sub-totals for each of the surfaces.

In Fig. 5.11**c** the two surfaces have joined to form one single surface containing two local minima in the ΔE function. It is in this case that care must be taken to ensure that each of the final states is included in the integration exactly once. Fig. 5.12 shows how a problem can arise.

In Fig. 5.12**a** the pair of joined surfaces is treated as a single surface. The result is that some of the surface lies 'in shadow' from the chosen origin, and is therefore not included in the integration.

In Fig. 5.12**b** the pair of joined surfaces are treated as separate, each being integrated over from its own origin. The result now is that parts of each surface are included in the integration twice.

In order to ensure surfaces are included just once, the integration is performed separately from each origin, as in Fig. 5.12**b**, but with the condition that if the gradient of ΔE with respect to a change in k in the radial direction is not always positive along the length of the radius, the integrand for this direction is taken to be zero. Fig 5.13

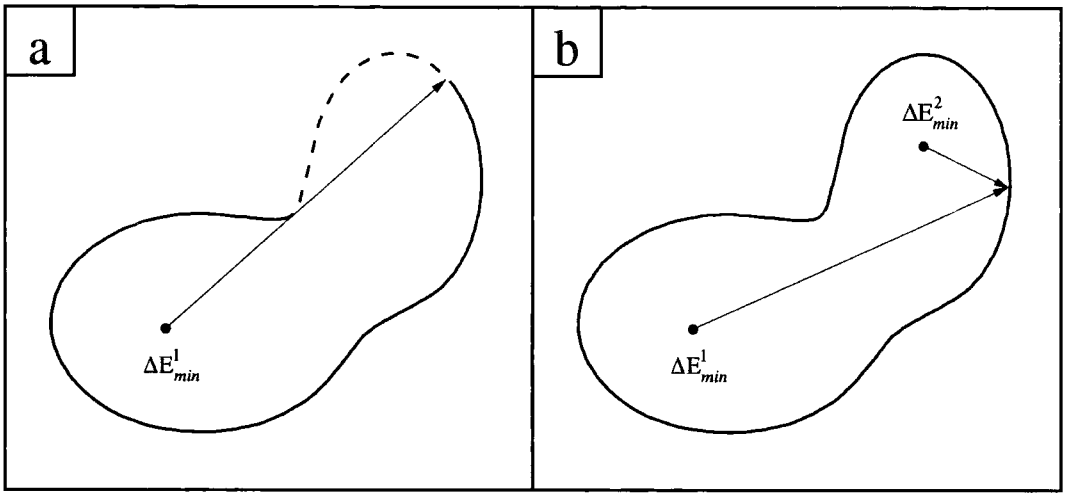


Figure 5.12: Problems of integrating over a complicated surface. In Fig. **a** the surface is integrated over from one origin, which results in some of the surface which lies 'in shadow' not being included. In Fig. **b**, the surface is integrated over from each of its minima, resulting in parts being counted twice.

illustrates how applying this condition leads to correct inclusion of all parts of the surface.

Fig. 5.13**a** shows a case similar to Fig. 5.11**b** in which there are two simple surfaces. Here integration takes place separately from origins ΔE^1_{min} and ΔE^2_{min}. The parts of the surface marked **A** and **B** on the figure are integrated from ΔE^1_{min} and the parts marked **C** and **D** are integrated from ΔE^2_{min}.

In Fig. 5.13**b** the two surfaces have merged to form a single one, which corresponds to the case shown in Fig. 5.11**c**. In this case we must be careful not to include the parts of the surface marked **E** and **F** twice, once from each origin. To avoid this, we stipulate that if the gradient of ΔE as a function of $(\mathbf{k}_{1'},\mathbf{k}_{2'})$ becomes negative as we move away from the origin, the corresponding integrand is taken to be zero. Thus, from ΔE^1_{min}, **E** is included, but **F** is taken to have zero integrand due to the gradient going negative at **G**. Similarly, **F** is included from ΔE^2_{min}, but **E** is not, again due to the gradient becoming negative at **G**. Hence **E** and **F** are included just once each during the integration.

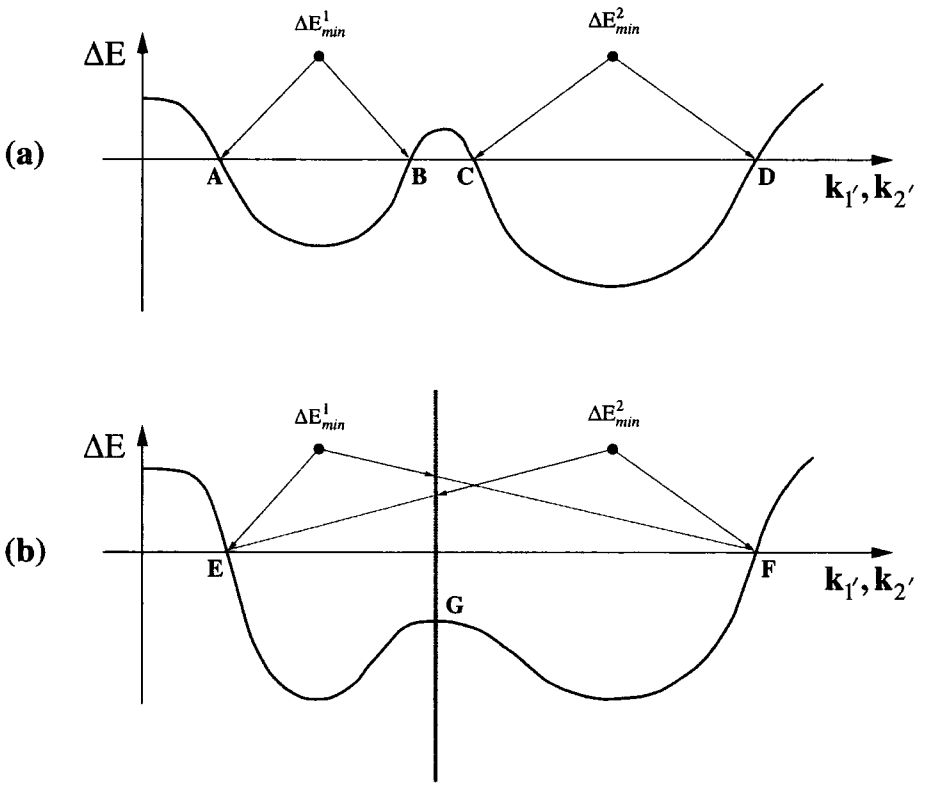


Figure 5.13: Solving the problem of integrating over complicated surfaces. Fig. **a** shows the situation represented in Fig. 5.11**b**: two simple surfaces exist, each integrated over separately. Fig. **b** shows the case represented in Fig. 5.11**c**: one complicated surface.

5.3.3 Performance of the Surface Algorithm

As discussed in §5.1.6, there are several numerical parameters used in the volume algorithm, specifically δe, B, N_{max} and N_{samp}, which must be adjusted to give convergent results. In the surface algorithm described here, there is only one parameter: the number of random sampling points taken, N_{samp}. This value should be set high enough to reduce the statistical error on the result to an acceptable level. In this work, sampling points are taken until the statistical error is 1% or below.

5.4 Comparison of Integration Methods

It is useful to compare the results of integrations performed using the different algorithms. They approach the problem in quite different ways, and therefore agreement in their results to within some small error can be taken as confirmation of a correct evaluation of the integral. However, it should be remembered that both algorithms use the same band structure data (discussed in Chapter 3) and the same method of calculating the matrix element (discussed in Chapter 4), and so it does not follow that the overall error on each of the results is of the same order as any discrepancy between them.

Fig. 5.14 shows rates calculated using the volume and surface integration methods for impacting electrons in the first conduction band of GaAs lying along the line defined by $\mathbf{k} = (0.32+t,\ 0.32-t,\ 0)$. For these states, only one surface of allowed transitions is accessible, corresponding a minimum in ΔE located near $\mathbf{k}_{1'} = \mathbf{k}_{2'} = \Gamma$ (i.e. the situation is as shown in Fig. 5.11**a**). Agreement is generally good, except where the rate is at its lowest in which case the volume algorithm becomes inaccurate.

Fig. 5.15 compares the results of the volume and surface algorithms, this time for impacting electrons located along the line $\mathbf{k} = (t,\ 0.055 + \frac{t}{2},\ 0)$ in the first conduction band of GaAs. For $0.3 \lesssim t \lesssim 0.45$ there is only one surface of allowed transitions accessible, which is centred near $\mathbf{k}_{1'} = \mathbf{k}_{2'} = \Gamma$, and the rates obtained using the volume

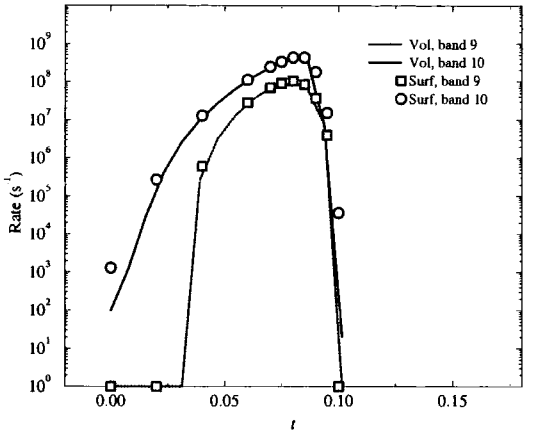


Figure 5.14: Comparison of rates in the 1st conduction band of GaAs obtained by the volume and surface algorithms when only one surface of allowed transitions is accessible.

and surface algorithms agree. For $t \gtrsim 0.45$, other surfaces corresponding to minima in ΔE located in the satellite valleys become available (i.e. the situation is as shown in Fig. 5.11**b**). The volume algorithm includes these other surfaces automatically, and the rate obtained by it correspondingly increases. The surface algorithm, which in this case is applied only to the minimum in ΔE at $\mathbf{k}_{1'} = \mathbf{k}_{2'} = \Gamma$, misses these new surfaces and therefore underestimates the rate when $t \gtrsim 0.45$.

Fig. 5.16 shows the rates plotted for the same impacting electrons as in Fig. 5.15, but this time the surface algorithm has been applied to all the accessible surfaces and the integrands summed. In this case there is good agreement between the rates calculated by the volume and surface algorithms.

5.4.1 Umklapp Processes

For any electron state at $\mathbf{k}$ in the first Brillouin zone there are equivalent states outside the zone at $\mathbf{k}' = \mathbf{k} + \mathbf{G}$ (where $\mathbf{G}$ is any reciprocal lattice vector) with the same energy and wavefunction. If the original state $\mathbf{k}$ lies on a surface of allowed (energy and momentum conserving) transitions, then any equivalent state $\mathbf{k}'$ lies on another

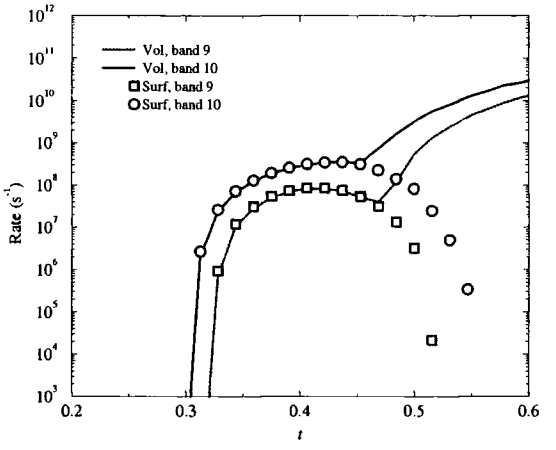


Figure 5.15: Comparison of rates in the 1st conduction band of GaAs obtained by the volume and surface algorithms when multiple surfaces of allowed transitions are available. The surface algorithm is only applied to the surface corresponding to the minimum in ΔE near $\mathbf{k}_{1'} = \mathbf{k}_{2'} = \Gamma$.

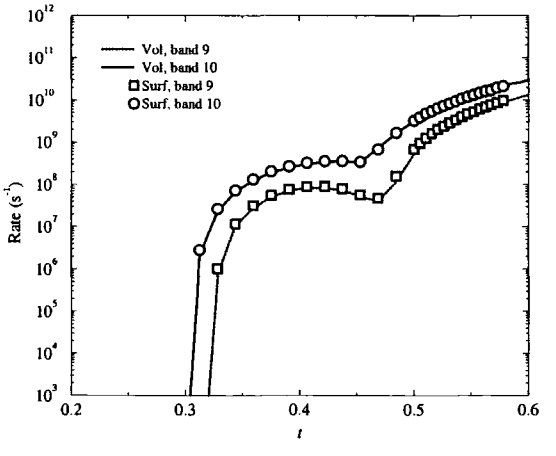


Figure 5.16: Comparison of rates in the 1st conduction band of GaAs obtained by the volume and surface algorithms when multiple surfaces of allowed transitions are available. The surface algorithm is applied to surfaces corresponding to all minima in ΔE.

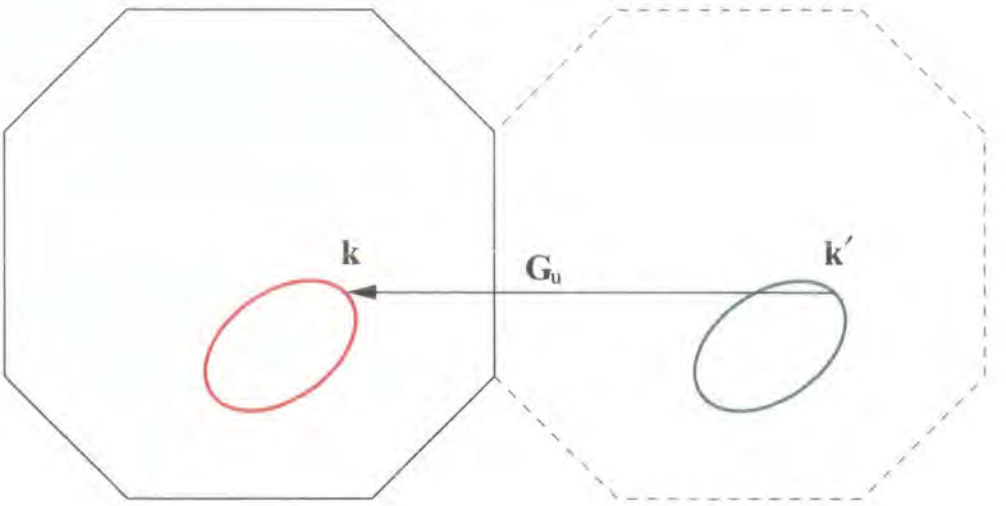


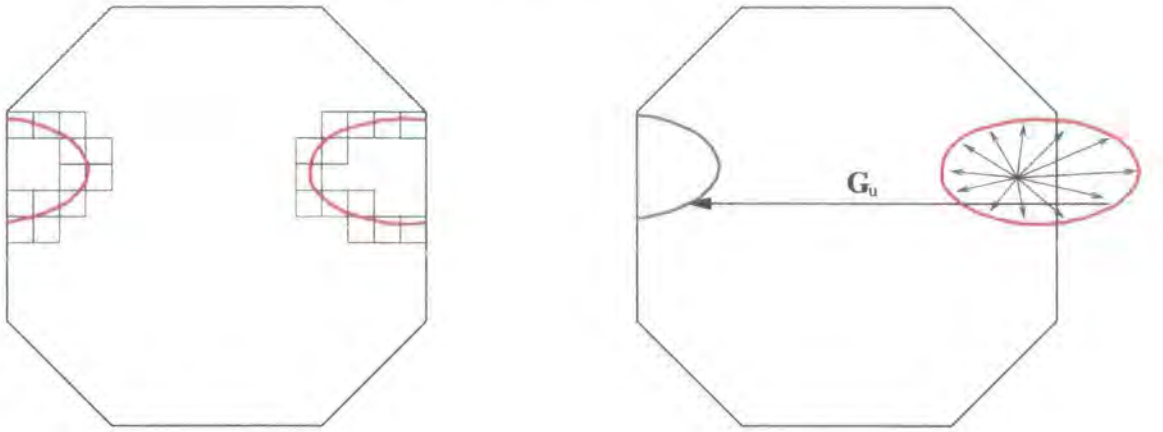
Figure 5.17: A schematic representation of equivalent surfaces of allowed (energy and momentum conserving) transitions, related by a reciprocal lattice vector. All states $\mathbf{k}'$ on the green surface are quantum mechanically equivalent to states $\mathbf{k}$ on the red surface, and only one surface should be included in the integration (usually the one in the first Brillouin zone).

surface of identical transitions. The situation is represented in Fig. 5.17.

In fact, the wavevectors $\mathbf{k}$ and $\mathbf{k}'$ just provide different descriptions of the same quantum mechanical state, and if one 'state' is included in the rate integral, the other should not be. That is, only one of the surfaces shown in Fig. 5.17 should be integrated over. In the case shown where the red surface is continuous within the first Brillouin zone it is sensible to work in the spirit of the reduced zone scheme and integrate over the surface in the first zone. However, if the surface is not continuous as in Fig. 5.18**a** then the volume and surface algorithms proceed in different ways.

The volume algorithm restricts all final state vectors to the first Brillouin zone, thus ensuring that no two transitions related by a reciprocal lattice vector are both included in the rate. This is represented in Fig. 5.18**a**. The $\mathbf{k}$-vectors involved in the transition conserve crystal momentum only to within a reciprocal lattice or umklapp vector, i.e. $\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_{1'} + \mathbf{k}_{2'} + \mathbf{G}_u$.

The surface algorithm, on the other hand, chooses the surface of allowed transitions in such a way as to be continuous as is represented in Fig. 5.18**b**. This is more conve-



(a) Volume algorithm (b) Surface algorithm

Figure 5.18: Different (but equivalent) surfaces of allowed transitions integrated over by the volume and surface algorithms. The volume algorithm requires that all final states lie in the first Brillouin zone, while the surface algorithm requires that the surface be continuous. In Fig. **b**, the part of the surface lying outside the zone is equivalent to the green surface lying inside the zone.

nient for the surface algorithm, which relies on choosing a coordinate origin *inside* the surface (see §5.3.1). Such a choice may lead to the final state **k**-vectors lying outside the first Brillouin zone. However the origin (or origins, in the case of there being multiple surfaces) from which the integration is performed, is restricted to the first Brillouin zone to avoid wrongly including transitions related by a reciprocal lattice vector.

5.5 Summation of Rates Over Band Index

Consider the transition rate from a state in band 12 due to its impact ionisation of states in band 8, leaving the electrons in bands 9 or 10 (i.e. initial states in the second conduction and heavy hole bands, final states in the first[c] conduction band (see Chapter 2, Table 2.2 for notation). There are four separate rates to be calculated, corresponding to the four possible combinations of final state bands. If transitions from the impacting band A and impacted band B to the final bands C and D are denoted

[c]In principle, either particle could make a transition to a higher conduction band, if energy and momentum allow, but transitions of this sort are neglected for the moment.

symbolically as A,B→C,D, then the four rate integrations correspond to transitions of the types:

12,08→09,09 12,08→09,10

12,08→10,09 12,08→10,10

The total rate of impact ionisation might then be assumed to be equal to the sum of rates of all the possible sub-transitions listed above. In fact, this is *not* the case, and instead summing the rates of all possible transitions gives a value for the total rate that is twice the actual theoretical value [64,107]. This is because in summing the rates due to all possible transitions, as described above, each *quantum-mechanically distinct* transition is counted twice. The two individual transitions

$$\mathbf{k}_1, \mathbf{k}_2 \rightarrow \mathbf{k}_{1'}, \mathbf{k}_{2'} \qquad \text{and} \qquad \mathbf{k}_1, \mathbf{k}_2 \rightarrow \mathbf{k}_{2'}, \mathbf{k}_{1'}$$

are not quantum-mechanically distinct because the second is the exchange of the first. They are therefore the same, and their contribution to the transition rate should only be counted once[d]. Note that the effect of exchanging the final states is accounted for explicitly in the matrix element, which consists of direct and exchange parts (see Chapter 4, §4.2).

Therefore, when integrating over all final pairs of **k**-vectors for transitions of the type 12,08→09,09, the value obtained for the rate must be halved to obtain the true theoretical value. The same is true of the rate obtained for 12,08→10,10.

In the case of the rate 12,08→09,10, simply swapping the final state **k**-vectors does not give the exchange transition since the band indices also differ. Thus all transitions included in the integral are quantum-mechanically distinct as they should be. However each transition in the rate 12,08→10,09 is the exchange process of a transition already summed over when integrating 12,08→09,10, and so including these again leads to double counting.

[d] Which one is counted is unimportant since the magnitude of the matrix element is the same in each case.

Thus the total rate of transitions from a state in band 12 due to impact ionisation of states in band 8 is given by

$$12,08\rightarrow CB1,CB1 = \frac{1}{2}\Big(12,08\rightarrow 09,09 + 12,08\rightarrow 09,10 + 12,08\rightarrow 10,09 + 12,08\rightarrow 10,10\Big) \quad (5.21)$$

$$= \frac{1}{2}(12,08\rightarrow 09,09) + 12,08\rightarrow 09,10 + \frac{1}{2}(12,08\rightarrow 10,10) \quad (5.22)$$

where the second form, which makes use of the relation $12,08\rightarrow 09,10 = 12,08\rightarrow 10,09$, is more convenient as it only requires three, rather than four, numerical integrations.

In this example, the rate obtained from Eq. (5.22) is that due to all transitions involving impacted carriers in band 8 and final states in the first conduction band. Generally the quantity of interest is the *total* rate of impact ionisation caused by a particular impacting carrier due to any allowed transitions. This total transition rate is obtained by summing the contributions from all valence and conduction bands for which energy and momentum conserving transitions exist.

5.5.1 Pros and Cons of the Two Algorithms

The two algorithms have different strengths and weaknesses which in many cases are complementary. The columns below list the main properties of each algorithm. Comments marked with a ⊕ correspond to advantageous characteristics, with those marked with a ⊖ corresponding to disadvantageous characteristics.

Volume Algorithm	Surface Algorithm
⊖ Impacting carrier wavevectors must lie at the nodes of the finest grid G_B on which energy maxima and minima are stored (see the discussion of pre-calculation of maxima and minima in §5.1.4).	⊕ Impacting carrier wavevectors are free to lie anywhere.
⊖ Care must be taken to ensure the numerical parameters δe, B, N_{max} and N_{samp} are chosen so that convergent rates are obtained.	⊕ The algorithm relies on only one parameter, N_{samp}, which can be chosen relatively easily to give an acceptable statistical error.
⊕ All transitions are integrated automatically in one pass of the algorithm.	⊖ The algorithm must be run separately for each sub-surface accessible to the impacting electron.
⊕ No matter how complicated the surface of final states, the algorithm sums over all transitions without special provisions being made.	⊖ As the surface of final states becomes complicated, extra care must be taken to avoid mis-counting the transitions to this surface (see §5.3.2).

Chapter 6

General Results

In this chapter, general results relating to the band structure, thresholds and rates of the materials studied are presented. The aim of the chapter is to provide a comprehensive survey of the results of the calculations described in the previous chapters. The results are discussed where appropriate, and are compared with calculations performed by other authors.

6.1 Terminology Used in this Chapter

In order to improve the clarity of the following sections, the definitions of some of the terms used in the rest of this chapter are given below.

Carrier is used in the usual sense to mean an electron in the conduction band or a hole in the valence band.

Carrier Energy is used to denote the energy of an electron above the conduction band edge, or a hole below the valence band edge. Thus for electrons, carrier energy increases with increasing energy eigenvalue of the occupied eigenstate, whereas for holes, carrier energy increases with decreasing energy eigenvalue of the unoccupied electron eigenstate.

State. Strictly, a state can be either occupied or unoccupied only by an electron. However, in the case of hole initiated impact ionisation, the hole will be discussed as if it were a real positively charged particle (as is conventional), and as such will be said to occupy a state. Additionally, for the sake of conciseness, expressions of the form 'the ionising state' (or similar) will be used when strictly what is meant is 'the state occupied by the ionising carrier'.

Secondary States are the impacted and final states. After an electron initiated impact ionisation has taken place, the final states are occupied states in the conduction band and the impacted state is a state unoccupied by an electron in the valence band (corresponding to a hole — see *Generated Carriers* below). After a hole initiated impact ionisation, the final states in the valence band are occupied by holes, while the impacted state in the conduction band is unoccupied by a hole (corresponding to an occupied electronic state — again, see below).

Generated Carriers are the three carriers remaining after a single impact ionisation event — two of the same type (electron or hole) as the ionising carrier, one of the opposite type. It should be noted that there is a distinction between the generated carriers and the secondary states. In the case of electron initiated impact ionisation, the generated electrons occupy the final secondary states. However, if the impacted secondary state (that is, the state left unoccupied by an electron) lies at $\mathbf{k}_2$, then the generated hole lies at $-\mathbf{k}_2$ (see, for example, Kittel [101]). Similarly, for hole initiated impact ionisation, the generated holes 'occupy' the final states, while the generated electron lies at minus the wavevector of the impacted state. Making this distinction will only be important in §6.5.1 where the $\mathbf{k}$-space distribution of secondary states in discussed.

Finally, throughout this chapter, where InGaAs and SiGe are referred to, the compositions are always $In_{0.53}Ga_{0.47}As$ and $Si_{0.5}Ge_{0.5}$, and all material parameters are for unstrained bulk material at 300K.

6.2 Band Structure

A summary of the parameters required to perform the pseudopotential band structure calculation (i.e. those listed in Chapter 2, Table 2.1) for each material is presented here, along with the results of the calculations.

GaAs

The pseudopotential parameters for GaAs were taken from Chelikowsky and Cohen [81]. The band structure calculation for GaAs involved several small differences in comparison to the calculations for InGaAs and SiGe[a]. Firstly, Gaussian non-local wells of the form $V_l = A_l \exp(-\frac{r^2}{R_l^2})$ were used instead of square wells as for the other materials. Secondly, the calculation of the spin orbit interaction included the $B_{nl}(K)$ terms which are neglected in the other materials studied here, as discussed in §2.3 of Chapter 2. Finally, the pseudopotential parameters used were for GaAs at 0K. When calculating band structure at 300K, as was used in all the calculations presented in this work, the band gap was 'scissored', i.e. conduction band energy eigenvalues were all reduced by 118 meV.

The values of the parameters used in the calculation are listed in Table 6.1, and energies of some of the important gaps obtained with these values are given in Table 6.2. The energy band structure and dielectric function resulting from the pseudopotential calculation are shown in Figs. 6.1 and 6.2.

InGaAs

An initial set of pseudopotential parameters were generated for InGaAs by interpolating parameters between the binary compounds, taken from Chelikowsky and Cohen [81].

[a]This is simply due to that fact that the pseudopotentials for InGaAs and SiGe were fitted for this work, whilst that for GaAs was already available.

The form factors were interpolated using the expression

$$V_{\text{alloy}}(\mathbf{G}) = \frac{1}{\Omega_{\text{alloy}}}\left[x\Omega_A V_A(\mathbf{G}) + (1-x)\Omega_B V_B(\mathbf{G})\right] \tag{6.1}$$

where Ω is the volume of the unit cell of the material indicated by the subscript, A and B refer to InAs and GaAs respectively, and $x = 0.53$. The non-local well depth parameters were interpolated linearly. This required the band structure for GaAs to be re-fitted, using a square non-local well, so that the pseudopotentials used for it and InAs would be of the same type. The initial estimate of the parameters thus obtained was then refined using the fitting procedure described in Chapter 3, §3.1. The experimental data for unstrained InGaAs at 300K used for fitting was taken from [75]. The final fitted pseudopotential parameters are listed in Table 6.3. Table 6.4 gives energy gaps calculated using these parameters along with the corresponding experimental data used for fitting. The energy band structure and dielectric function obtained from the pseudopotential calculation are shown in Figs 6.3 and 6.4.

SiGe

Pseudopotential parameters for SiGe were obtained in a similar way to those of InGaAs. Using local pseudopotential form factors for Si and Ge given in [110], a local pseudopotential for SiGe was obtained by interpolation with Eq. (6.1). The interpolated form factors were then used as the starting point for the fit to experimental data. The data for unstrained SiGe at 300K was taken from [98]. The final fitted pseudopotential parameters are listed in Table 6.5 with energy gaps thus obtained listed in Table 6.6, along with the corresponding experimental values. The energy band structure and dielectric function are shown in Figs.6.5 and 6.6.

Parameter	Value	Parameter	Value
$V_S(\sqrt{3})$	-0.2140	α_0^c	0.000
$V_S(\sqrt{8})$	0.0140	β_0^c	0.000
$V_S(\sqrt{11})$	0.0670	A_2^c	0.125
$V_A(\sqrt{3})$	0.0550	R_0^c	1.296
$V_A(\sqrt{4})$	0.0380	R_2^c	1.219
$V_A(\sqrt{11})$	0.0010	α_0^a	0.000
		β_0^a	0.000
α	1.38	A_2^a	0.625
a_0	5.648	R_0^a	1.058
		R_2^a	1.219

Table 6.1: The pseudopotential parameters for GaAs (see Chapter 2, Table 2.1 for the meanings of the symbols). V_A, V_S, α_0, and A_2 are in units of Rydbergs; R and a_0 are in Å; β_0 and α are dimensionless. μ was adjusted to give the $\Gamma_{\text{HH}} - \Gamma_{\text{SSO}}$ splitting listed in Table 6.2.

Energy gap	Pseudopotential value
$\Gamma_{\text{CB1}} - \Gamma_{\text{HH}}$ (E_g)	1.540†
$X_{\text{CB1}} - \Gamma_{\text{CB1}}$	0.501
$L_{\text{CB1}} - \Gamma_{\text{CB1}}$	0.311
$\Gamma_{\text{HH}} - \Gamma_{\text{SSO}}$	0.350
Position of X	(0.86, 0, 0)
Position of L	(0.50, 0.50, 0.50)

Table 6.2: Energy gaps in GaAs (in eV), obtained from the pseudopotential calculation. P_B denotes the energy eigenvalue (as opposed to carrier energy) at position P in band B. The letters X and L denote the positions of the minima in the 1st conduction band, *not* the symmetry positions at the zone boundary.

†In all calculations, the band gap is 'scissored' to 1.422 eV

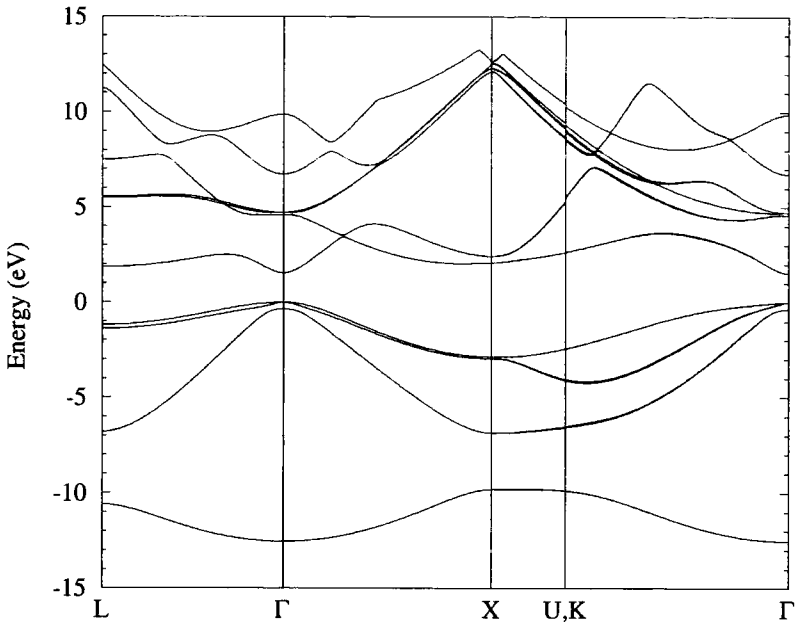


Figure 6.1: The first 20 energy bands of GaAs obtained from the pseudopotential parameters listed in Table 6.1.

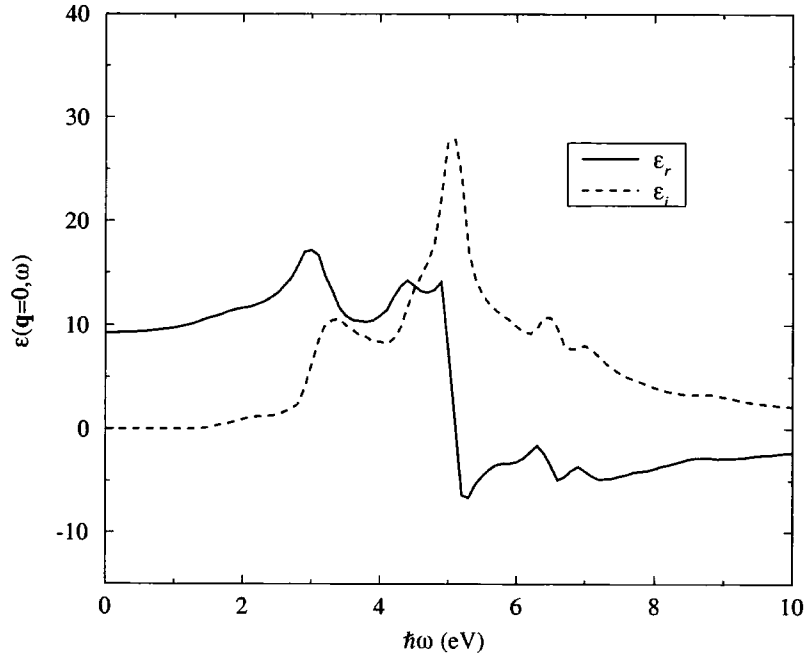


Figure 6.2: The dielectric function of GaAs for $\mathbf{q} = 0$, obtained from the band structure shown in Fig. 6.1. The real part is shown as the solid line, the imaginary part as the dashed line.

Parameter	Value	Parameter	Value
$V_S(\sqrt{3})$	-0.2064	α_0^c	0.0000
$V_S(\sqrt{8})$	0.0065	β_0^c	0.0005
$V_S(\sqrt{11})$	0.0558	A_2^c	0.5575
$V_A(\sqrt{3})$	0.0480	R_0^c	1.2696
$V_A(\sqrt{4})$	0.0441	R_2^c	1.2691
$V_A(\sqrt{11})$	0.0092	α_0^a	0.0000
		β_0^a	0.1287
α	0.9927	A_2^a	1.5583
a_0	5.8618	R_0^a	1.0580
		R_2^a	1.2691

Table 6.3: The pseudopotential parameters for InGaAs (see Chapter 2, Table 2.1 for the meanings of the symbols). V_A, V_S, α_0, and A_2 are in units of Rydbergs; R and a_0 are in Å; β_0 and α are dimensionless. μ was adjusted to give the $\Gamma_{\text{HH}} - \Gamma_{\text{SSO}}$ splitting listed in Table 6.4.

Energy gap	Pseudopotential value	Experimental value
$\Gamma_{\text{CB1}} - \Gamma_{\text{HH}}$ (E_g)	0.749	0.75
$X_{\text{CB1}} - \Gamma_{\text{CB1}}$	0.671	0.67
$L_{\text{CB1}} - \Gamma_{\text{CB1}}$	0.552	0.55
$X_{\text{CB1}} - X_{\text{HH}}$	4.293	4.84
$L_{\text{CB1}} - L_{\text{HH}}$	2.528	2.55
$\Gamma_{\text{CB2}} - \Gamma_{\text{HH}}$	4.312	4.33
$\Gamma_{\text{HH}} - \Gamma_{\text{SSO}}$	0.360	0.35
$L_{\text{HH}} - L_{\text{LH}}$	0.203	0.27
Position of X		(0.99, 0, 0)
Position of L		(0.49, 0.49, 0.49)

Table 6.4: Energy gaps in InGaAs (in eV), obtained from the pseudopotential calculation. See also caption to Table. 6.2.

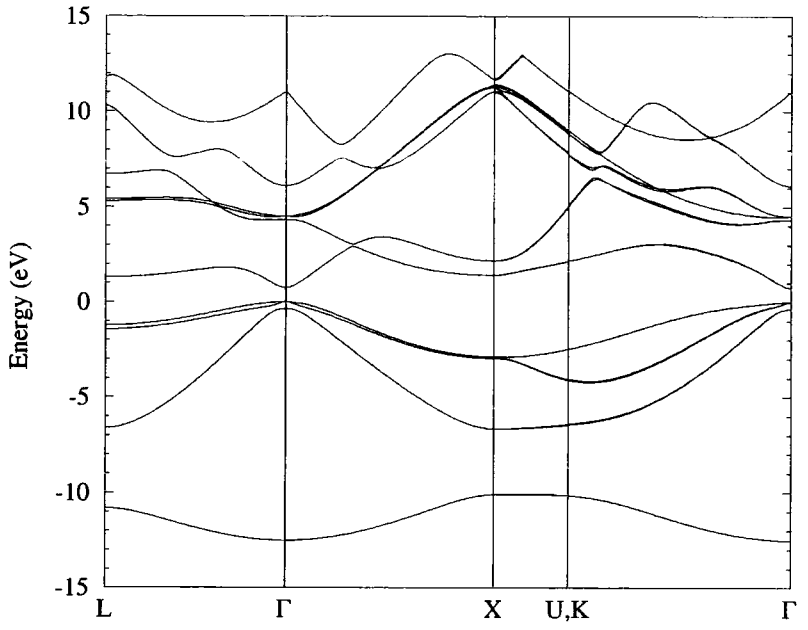


Figure 6.3: The first 20 energy bands of InGaAs obtained from the pseudopotential parameters listed in Table 6.3.

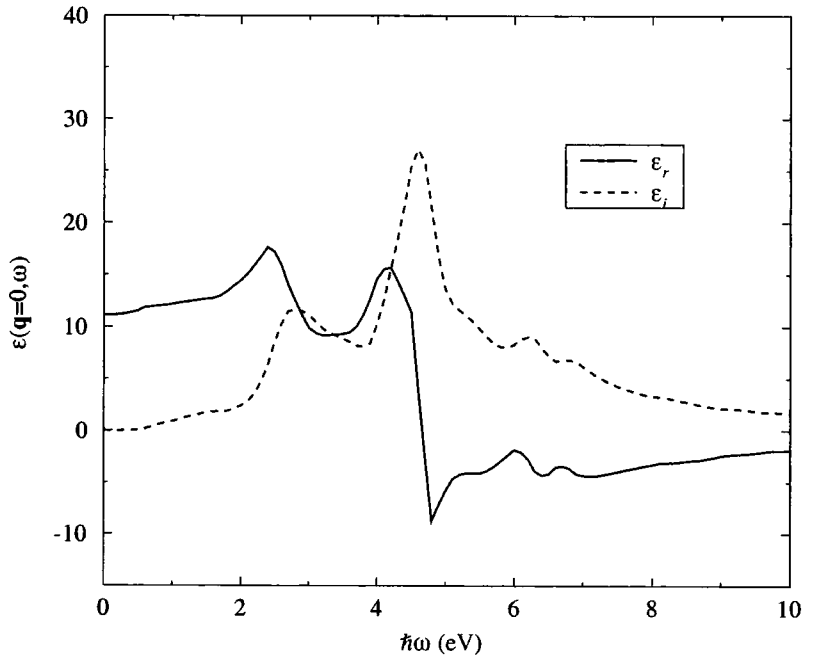


Figure 6.4: The dielectric function of InGaAs for $\mathbf{q} = 0$, obtained from the band structure shown in Fig. 6.3. The real part is shown as the solid line, the imaginary part as the dashed line.

Parameter	Value	Parameter	Value
$V_S(\sqrt{3})$	-0.225548	α_0^c	0.003569
$V_S(\sqrt{8})$	0.026800	β_0^c	0.200779
$V_S(\sqrt{11})$	0.064081	A_2^c	0.526179
$V_A(\sqrt{3})$	0.000000	R_0^c	1.059587
$V_A(\sqrt{4})$	0.000000	R_2^c	1.198185
$V_A(\sqrt{11})$	0.000000	α_0^a	0.003569
		β_0^a	0.200779
α	1.0	A_2^a	0.526179
a_0	5.5344	R_0^a	1.059587
		R_2^a	1.198185

Table 6.5: The pseudopotential parameters for SiGe (see Chapter 2, Table 2.1 for the meanings of the symbols). V_A, V_S, α_0, and A_2 are in units of Rydbergs; R and a_0 are in Å; β_0 and α are dimensionless. μ was adjusted to give the $\Gamma_{HH} - \Gamma_{SSO}$ splitting listed in Table 6.6.

Energy gap	Pseudopotential value	Experimental value
$X_{CB1} - \Gamma_{HH}$ (E_g)	0.908	0.91
$\Gamma_{CB1} - \Gamma_{HH}$	2.360	2.41
$X_{CB1} - X_{HH}$	4.194	4.36
$L_{CB1} - L_{HH}$	2.942	2.72
$L_{CB1} - L_{LH}$	3.010	2.87
$\Gamma_{CB2} - \Gamma_{SSO}$	3.265	3.24
$\Gamma_{HH} - \Gamma_{SSO}$	0.115	0.12
Position of X		(0.82, 0, 0)
Position of L		(0.49, 0.49, 0.49)

Table 6.6: Energy gaps in SiGe (in eV), obtained from the pseudopotential calculation. See also caption to Table. 6.2.

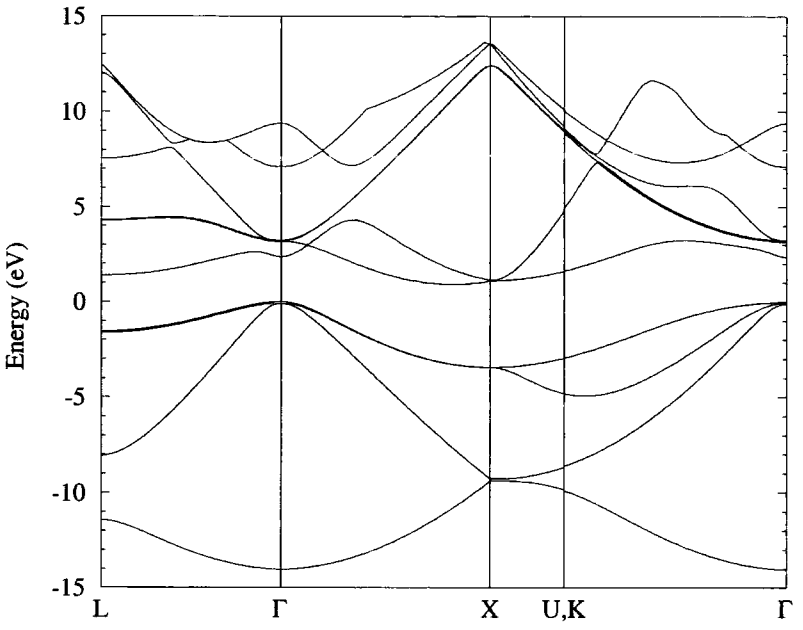


Figure 6.5: The first 20 energy bands of SiGe obtained from the pseudopotential parameters listed in Table 6.5.

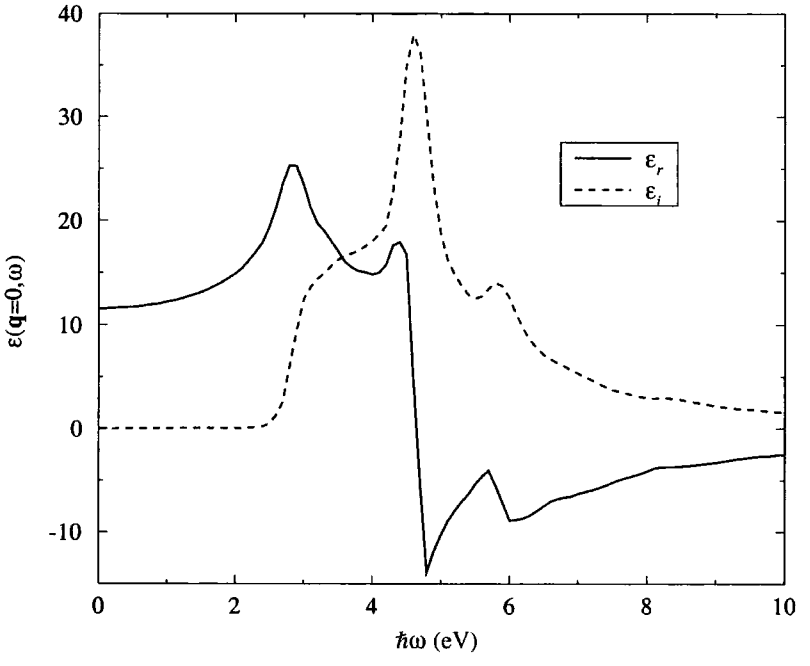


Figure 6.6: The dielectric function of SiGe for $\mathbf{q} = 0$, obtained from the band structure shown in Fig. 6.5. The real part is shown as the solid line, the imaginary part as the dashed line.

6.3 Impact Ionisation Thresholds

Impact ionisation thresholds are discussed in Chapter 4, §4.3. Although the thresholds give no information on the actual magnitude of the rate, they are important in determining the number of carriers in a device that can initiate impact ionisation and so it is of interest to determine where in **k**-space they lie. In this section the positions in **k**-space from which carriers can initiate ionisation, and the corresponding energies of these states are presented and compared between the different bands and materials studied.

6.3.1 Thresholds with respect to k-vector

The position in **k**-space of those states from which impact ionisation can be initiated was determined using the procedure described in §4.3.3. Briefly, the algorithm involves searching the final state phase space $(\mathbf{k}_{1'}, \mathbf{k}_{2'})$ for the minimum in the energy difference function ΔE (defined in §4.3.1 of Chapter 4) given a fixed value of the initiating carrier wavevector $\mathbf{k}_1$. If the minimum lies below zero, then a carrier in state $\mathbf{k}_1$ is able to initiate impact ionisation.

Values of ΔE_{min} were determined for impacting carriers in a given band whose wavevectors were located on a grid of points distributed throughout the irreducible wedge. The value of ΔE_{min} could then be interpolated at all points throughout the Brillouin zone in exactly the same way as band energies. Hence, by testing whether the interpolated values of the energy difference function lie above or below zero, the position of the thresholds can be located throughout the Brillouin zone.

The plots of the thresholds in Figs. 6.7–6.9 are all of the same type: the octagonal base of the plot is the $k_z = 0$ plane of the Brillouin zone with the energy of carriers[b] measured on the vertical axis. The energy surface is shaded dark in regions where

[b]Recall from §6.1 that the carrier energy of holes increases as the energy eigenvalue of the corresponding unoccupied electronic state decreases — hence the inversion of the valence band energy surfaces

impact ionisation can be initiated, light in regions where it cannot.

Thresholds in GaAs are shown in Fig. 6.7 for the first and second conduction bands and the heavy hole, light hole and spin split off bands. It is clear that in each band the main factor governing whether a carrier can initiate impact ionisation is its energy: the ionising states are those of higher energy. The restriction of impacting states to the higher energy regions of **k**-space results in the thresholds displaying anisotropy reflecting that of the energy bands themselves. Thus the spin split off band, which is the most isotropic, has the most isotropic threshold (i.e. for the 2-D plots shown here, the most circular). Other bands, particularly the conduction bands, have highly anisotropic thresholds.

Although the surface of threshold wavevectors is highly anisotropic, it corresponds to approximately constant energy. If the threshold were to lie exactly on an energy contour then whether a carrier were able to initiate impact ionisation would be determined by its energy alone (although the rate would in general vary with direction in **k**-space). In the materials studied here this is not the case, and in each band there exists a range of energies for which determination of a carrier's ability to initiate impact ionisation requires knowledge of its actual **k**-vector. This is discussed in §6.3.2.

Figs. 6.8 and 6.9 show the thresholds in the first conduction bands of InGaAs and SiGe. It can be seen that the thresholds in these materials behave in a similar way to GaAs, i.e. a carrier's energy is the main factor influencing its ability to impact ionise. The band structure of InGaAs is a very similar shape to that of GaAs, differing mainly in that the band gap of InGaAs is about half that of GaAs. The energy carriers must obtain is correspondingly lower and hence impact ionisation is possible from states throughout more of the zone. Thresholds are also of a similar form in SiGe, although impact ionisation is possible from all points within SiGe's shallow Γ-valley.

The high degree of anisotropy in the thresholds in **k**-space does not necessarily imply that similar anisotropy will be observed in the impact ionisation coefficients α and β for fields applied in different directions with respect to the crystallographic axes.

The anisotropy of the thresholds in **k**-space will only be reflected in α and β if ballistic electrons are primarily responsible for causing impact ionisation, as supposed by Shockley [46]. If, as suggested by other theories [17,45,47,48], carriers have undergone many phonon-scattering events before reaching threshold, they will be scattered throughout the zone. Thus **k**-space anisotropies will be 'integrated out' and the observed α and β coefficients will be isotropic with respect to field-crystal orientation.

The α and β coefficients are found to be isotropic in, for example, Si [25,31,38] and InP [39,40]. In GaAs, there is some disagreement over the anisotropy of the α coefficient. Experimentally, it has been observed to be anisotropic [41–43] while theoretically it is predicted to be isotropic [17]. However, both experimentally and theoretically, field directions were considered along the 100, 110 and 111 directions (and at various orientations between 110 and 111 in [17]). Along 111, no threshold can be reached ballistically, so for fields oriented in this direction, all electrons reaching threshold must have been scattered at least once. Along 100, electrons can only reach threshold ballistically by tunnelling into the second conduction band and along 110 a very small region of **k**-space from which ionisation can be initiated can be reached ballistically. Therefore, along all these directions we would expect the role of ballistic electrons to be limited in causing impact ionisation and hence the thresholds to be more-or-less isotropic. Examining the plot of thresholds for the first conduction band in GaAs in Fig. 6.7 it is clear that the majority of ionising states lie approximately along the 210 line. We expect therefore that if a ballistic contribution to the overall impact ionisation rate is to be measured, it would be for a field applied along this axis. To the author's knowledge, no such measurement has been made, which is presumably due to difficulties in growing crystals along such a direction.

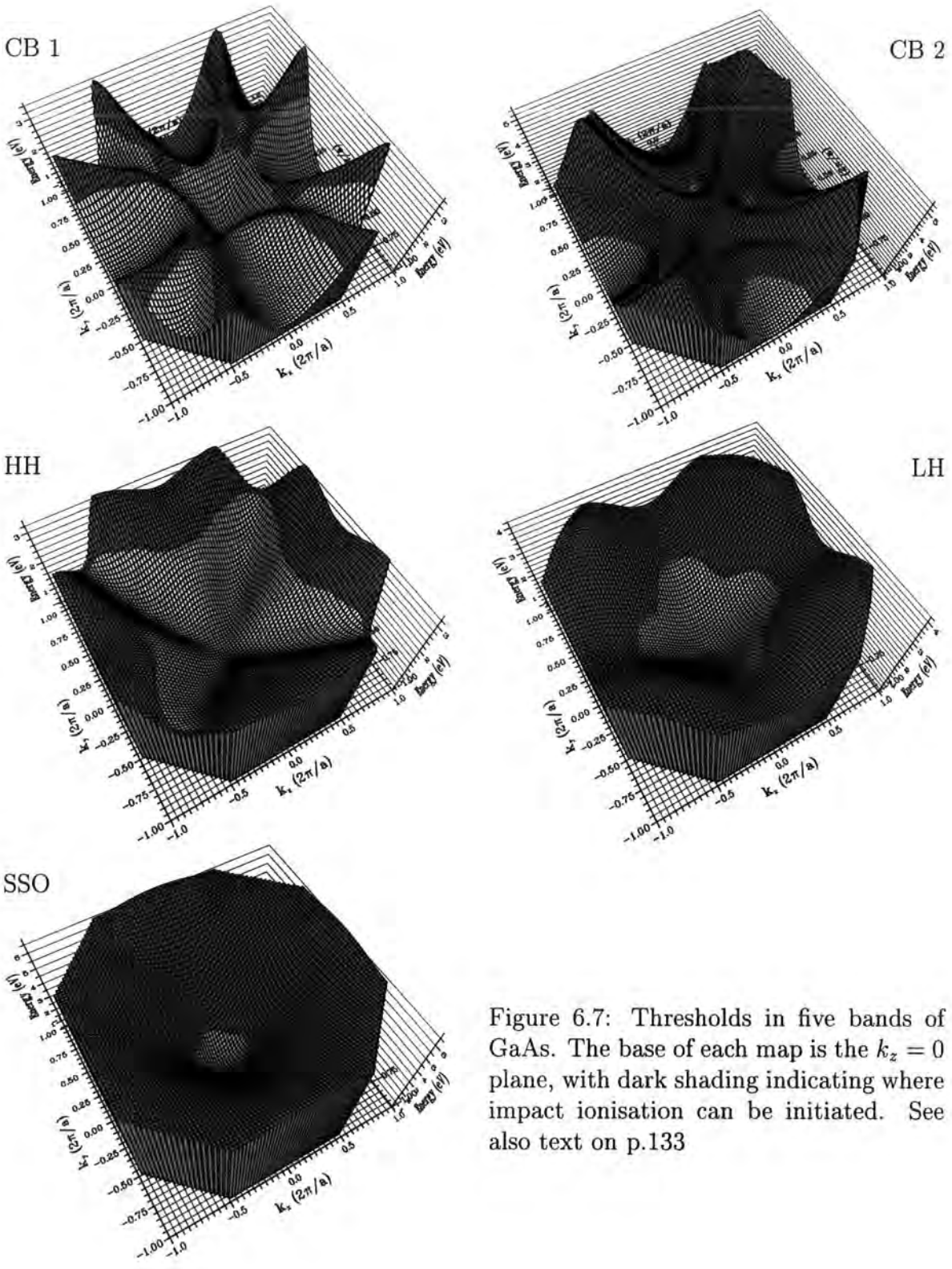


Figure 6.7: Thresholds in five bands of GaAs. The base of each map is the $k_z = 0$ plane, with dark shading indicating where impact ionisation can be initiated. See also text on p.133

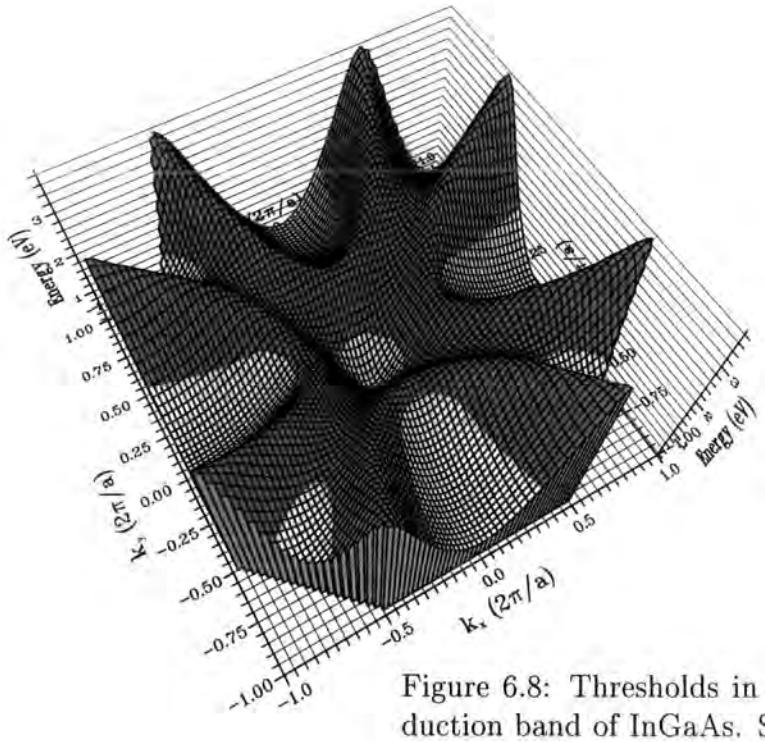


Figure 6.8: Thresholds in the 1st conduction band of InGaAs. See also text on p.133.

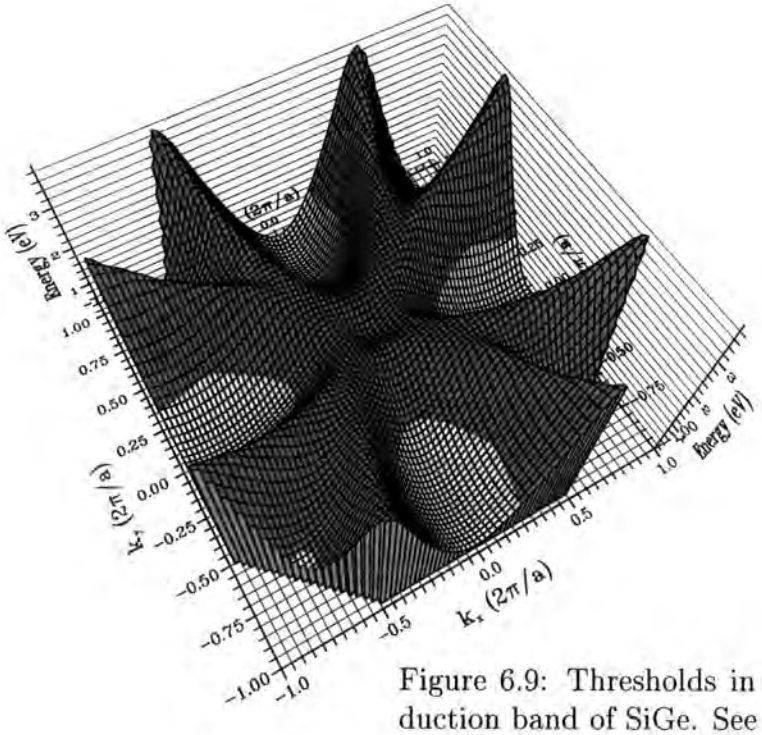


Figure 6.9: Thresholds in the 1st conduction band of SiGe. See also text on p.133.

6.3.2 Thresholds with respect to Energy

The thresholds presented in §6.3.1 as functions of the wavevector of the initiating carrier are reconsidered here as functions of the carrier energy. A fraction f is defined for a given band n as

$$f_n(E_i) = \frac{\int t_n(\mathbf{k})\ \delta(E_n(\mathbf{k}) - E_i)\ \mathrm{d}^3\mathbf{k}}{\int \delta(E_n(\mathbf{k}) - E_i)\ \mathrm{d}^3\mathbf{k}} \tag{6.2}$$

where E_i is impacting carrier energy, $E_n(\mathbf{k})$ is the carrier energy in band n at wavevector $\mathbf{k}$, and $t_n(\mathbf{k})$ is defined as a function whose value is 1 if state $\mathbf{k}$ in band n can initiate impact ionisation and zero otherwise. The integrals with respect to $\mathbf{k}$ are performed over the first Brillouin zone.

The denominator of Eq. (6.2) is proportional to the density of states at energy E_i, and the numerator is proportional to the density of states capable of initiating impact ionisation (the constant of proportionality being the same in each case). Thus $f_n(E_i)$ is the fraction of carriers at energy E_i which can initiate impact ionisation.

If, as mentioned in §6.3.1, the thresholds in $\mathbf{k}$-space were to lie along an energy contour, the function $f(E)$ would be a step function, rising from 0 to 1 at the energy of the contour on which the thresholds lay. The actual variation of $f(E)$ in each material is plotted in Figs. 6.10 to 6.12. As the plots show, the fraction of ionising states is not a step function but instead rises rapidly from 0 to 1 over an energy range of the order of 1 eV in most bands. Within this energy range a carrier's ability to initiate impact ionisation is influenced by the details of the band structure and is dependent on its actual $\mathbf{k}$-vector. Note that every state in the second conduction band of InGaAs can initiate impact ionisation and so $f(E)$ for this band is represented as a step function, rising from 0 to 1 at the energy of the bottom of the band.

In each material, the range of energy in which $f(E)$ rises from 0 to 1 is the greatest for the first conduction band. This reflects the fact that this is a band of complicated shape and low carrier energy, and hence simultaneously satisfying energy and momentum conservation is most difficult in this band. The thresholds therefore show the

greatest dependence on actual carrier wavevector, rather than just energy, particularly in InGaAs where the form of the $f(E)$ function is clearly influenced by the shape of the band.

The thresholds in the valence bands are similar for the direct gap materials GaAs and InGaAs. The absolute threshold for holes lies in the spin split off band, despite this band being of generally higher energy than the light and heavy hole bands. This reflects the fact that the difficulty of simultaneously conserving energy and momentum pushes the lowest ionising states to higher energy in the light and heavy hole bands. The spin split off bands also show the most step-like behaviour in $f(E)$. This can be understood by noting that all states involved in transitions near the threshold lie close to Γ (the locations of the secondary states involved in impact ionisation processes is discussed in §6.5) where anisotropy in all the relevant bands (of the impacting, impacted and final states) is least. Thus we expect to find the threshold at about the same $|\mathbf{k}|$ (and hence energy) in all directions.

The valence bands in the indirect gap SiGe show qualitatively different threshold behaviour to that of the direct gap materials. As well as the fact that the energy threshold for holes in SiGe lies above that for electrons, in contrast to the direct band gap materials, the rise of $f(E)$ from 0 to 1 occurs for the valence bands in the opposite order than in the direct gap case, i.e. the absolute threshold lies in the heavy hole band instead of the spin split off band. This is due to the fact that the lowest energy transition across the band gap corresponds to one of high $\mathbf{q}$ for indirect gap materials. The lowest energy states able to provide such momentum transfer lie in the heavy and light hole bands which have higher effective masses than the spin split off band. The indirect gap also accounts for the lack of step-like behaviour in the spin split off band. The arguments based on spherical symmetry made in the case of the direct gap materials do not apply to the indirect gap SiGe, and thus the isotopic behaviour of the threshold is not present.

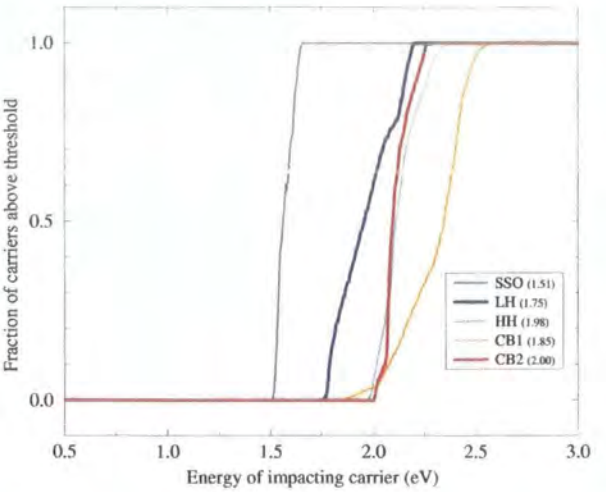


Figure 6.10: The fraction of ionising states at a given energy for each band in GaAs. The key indicates the energy threshold in each band.

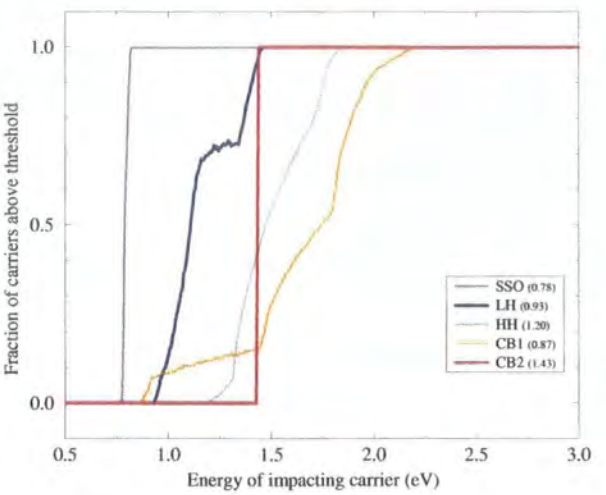


Figure 6.11: The fraction of ionising states at a given energy for each band in InGaAs.

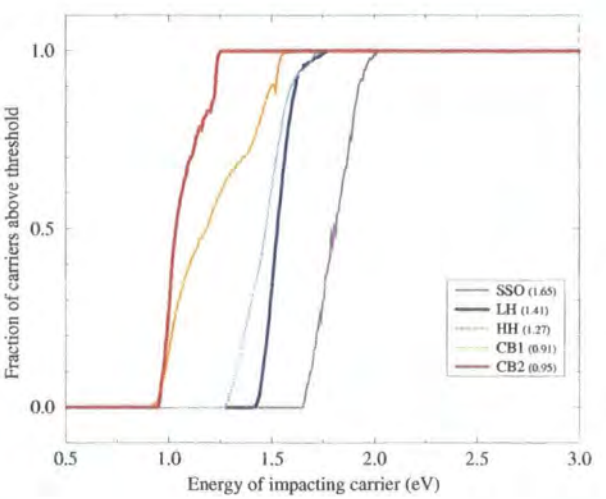


Figure 6.12: The fraction of ionising states at a given energy for each band in SiGe.

6.4 Impact Ionisation Rates

The impact ionisation rates presented in this and subsequent sections were calculated using the volume integration algorithm discussed in §5.1.2 of Chapter 5. For both electron and hole initiated processes, impacted and final states in the lower two conduction bands and upper three valence bands were included. To avoid excessive computational requirements, not all band combinations were included, but those neglected accounted for $\lesssim 5\%$ of the total rate.

The variation of the rate with respect to the wavevector and energy of the impacting carriers is examined below.

6.4.1 Rates with respect to k-vector

Rates in each band for each material are plotted here with respect to the impacting carrier's wavevector along the lines Γ–X, Γ–K and Γ–L (with the exception of the first conduction band of GaAs for reasons made clear below). All **k**-states in SiGe, and states lying along the Γ–X and Γ–L lines in GaAs and InGaAs are at least doubly spin degenerate. For such degenerate states, a single rate is plotted which is the average of the rates for each degenerate band. Along the Γ–K line in GaAs and InGaAs, the spin-orbit interaction splits the degeneracy and a pair of lines is plotted, one for each band.

The rates in the first conduction band of GaAs, InGaAs and SiGe are shown in Figs. 6.13–6.15. In GaAs it is not useful to plot along the X, K and L directions as they do not significantly intersect the small regions of **k**-space from which impact ionisation can be initiated, and instead rates are plotted along the four lines shown in Fig. 6.13. Rates in the second conduction bands of each material and the third conduction band of GaAs, plotted along Γ–X, Γ–K and Γ–L, are shown in Figs. 6.16–6.19.

The complicated variation of the rates with impacting **k**-vector reflects the complexity of the energy band structure in the conduction band, and the rates are in

general highly anisotropic functions. Several discontinuities in the value of the rate with respect to **k**-vector can be seen in the various plots: along Γ–X in the first and second conduction bands of SiGe, and along Γ–L in the second conduction band of InGaAs and second and third conduction bands of GaAs. These points correspond to crossing points in the energy bands, and the discontinuity in the rate is caused by a discontinuity in the matrix elements as the bands cross. (In the case of the volume of phase space, which is discussed in §7.1 of Chapter 5, the discontinuity is in the first derivative).

Where the rate is highest it shows qualitatively similar behaviour from material to material, while where it is low the materials differ considerably. Thus the plots along Γ–K and Γ–L in the second conduction band are similar in shape for all materials, while along Γ–X of the second conduction band and in all directions in the first conduction band, the rates are quite different.

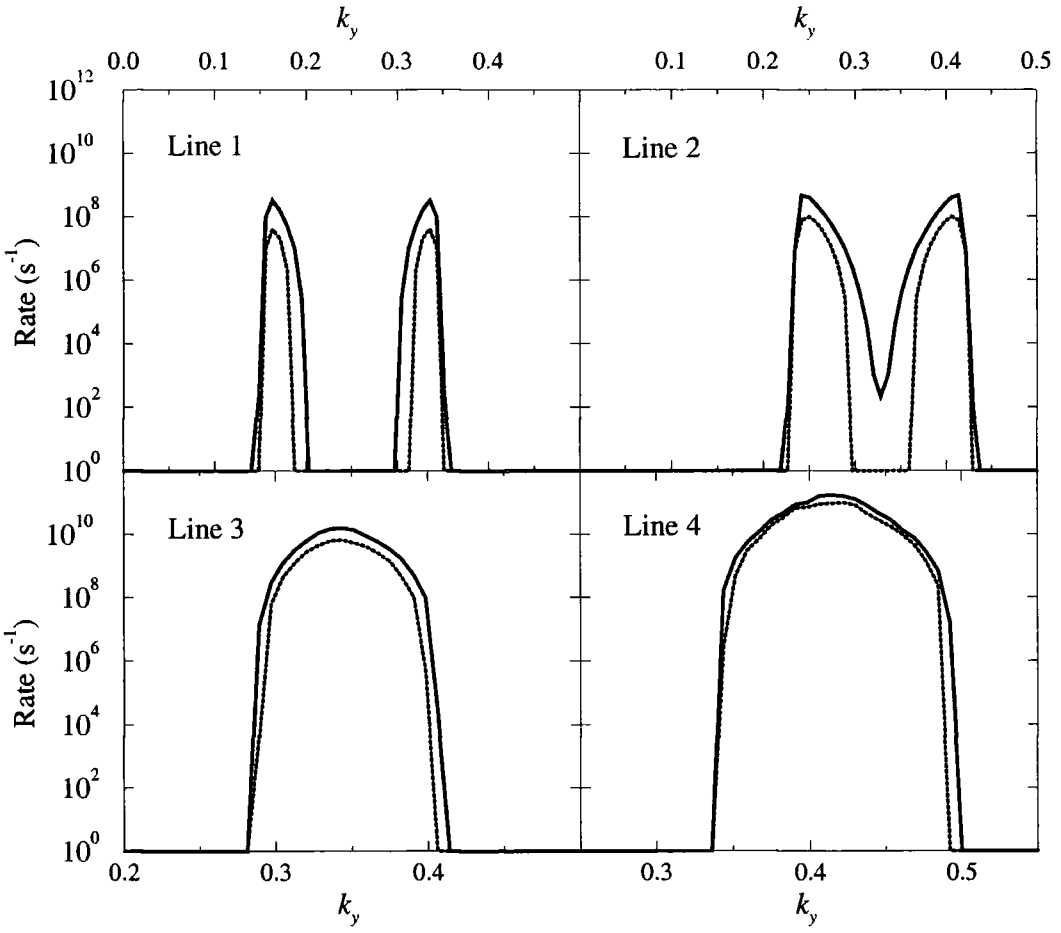
A high rate corresponds to a large surface or surfaces of final states. In such cases, small differences in the shape of the bands have little effect on the availability of allowed transitions, and hence the three materials, having broadly similar second conduction band structures, show similar behaviour in the rates. Where the rate is low it is due to the surfaces of allowed final states being small. In this case, subtle differences in band structure can radically alter the availability of final states and hence the rates differ considerably between materials.

As discussed in Chapter 2, the spin-orbit interaction breaks the double degeneracy of the bands in GaAs and InGaAs. The effect this has on the rates can be seen in the plots along the non-degenerate Γ–K line, and in all the lines plotted for the first conduction band of GaAs. It can be seen that there is a significant difference in the rates for the upper and lower spin-split bands only where the rate is lowest, i.e. in the first conduction band. Along Γ–K in the second conduction bands of these materials (and the third conduction band of GaAs) the splitting is of less significance.

Fig. 6.20 shows the rates in the first and second conduction bands of GaAs through-

out the $k_z = 0$ plane of the Brillouin zone. This figure can be compared to the threshold plots in Fig. 6.7. Obviously, the regions of non-zero rates match the dark-shaded regions denoting ionising states on the threshold plots. A qualitative correspondence between the energy band structure and the rate is also clear for each of the two bands.

Figs. 6.21–6.23 show the rates of hole initiated impact ionisation plotted along the Γ–X, Γ–K and Γ–L lines for each material. Variation of the rate with **k**-vector is of a simple monotonic form, reflecting the simpler structure of the valence bands themselves. As with the conduction bands, where the rate is highest the most similarity is seen between materials. Thus the spin split off bands of each material show quite good quantitative correspondence, while the light hole and particularly the heavy hole bands show less similarity.



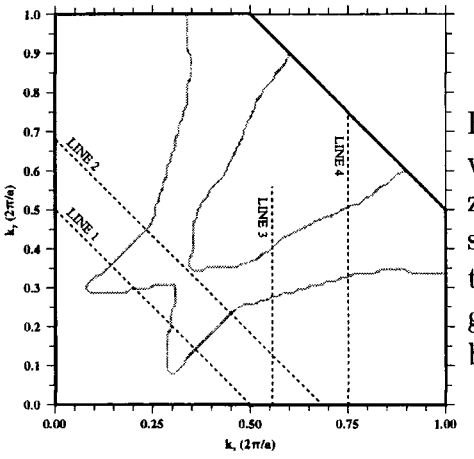


Figure 6.13: In the plot above, rates are shown with respect to k_y for four lines in the Brillouin zone. The dotted lines correspond to band 9, the solid lines to band 10. The plot to the left shows the four lines which all lie in the $k_z = 0$ plane. The grey line indicates the position of the threshold in band 10 within this plane.

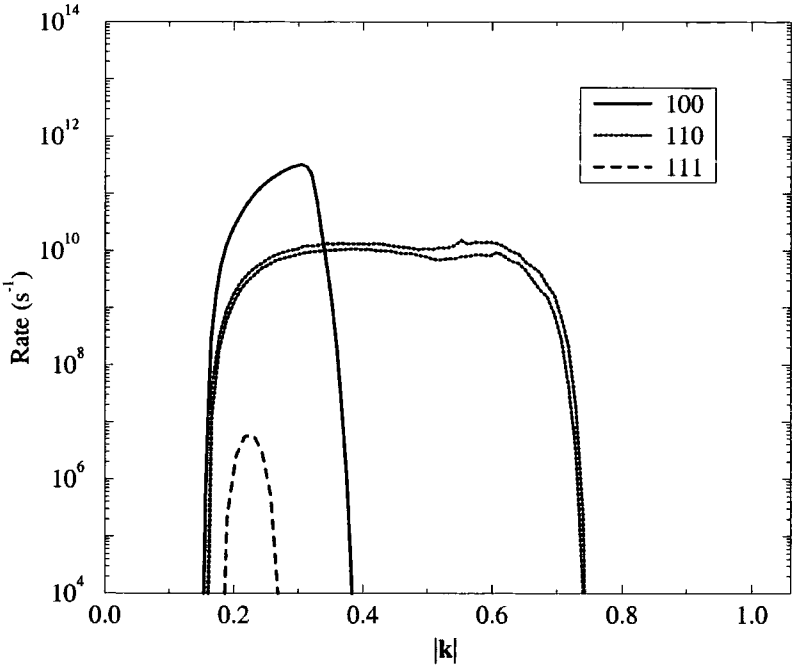


Figure 6.14: Rates in the 1st conduction band of InGaAs plotted with respect to **k**-vector of the initiating carrier along the lines Γ–X, Γ–K and Γ–L.

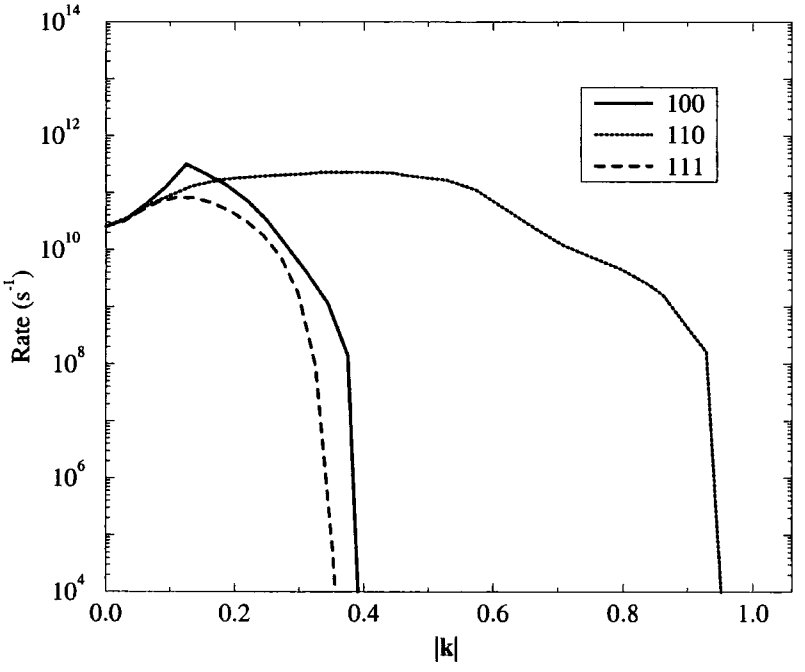


Figure 6.15: Rates in the 1st conduction band of SiGe plotted with respect to **k**-vector of the initiating carrier along the lines Γ–X, Γ–K and Γ–L.

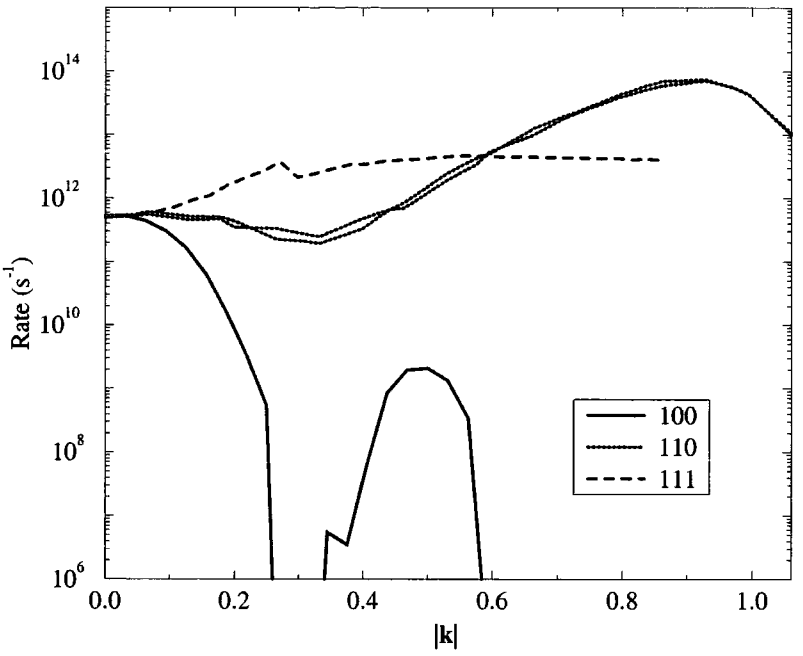


Figure 6.16: Rates in the 2nd conduction band of GaAs plotted with respect to **k**-vector of the initiating carrier along the lines Γ–X, Γ–K and Γ–L.

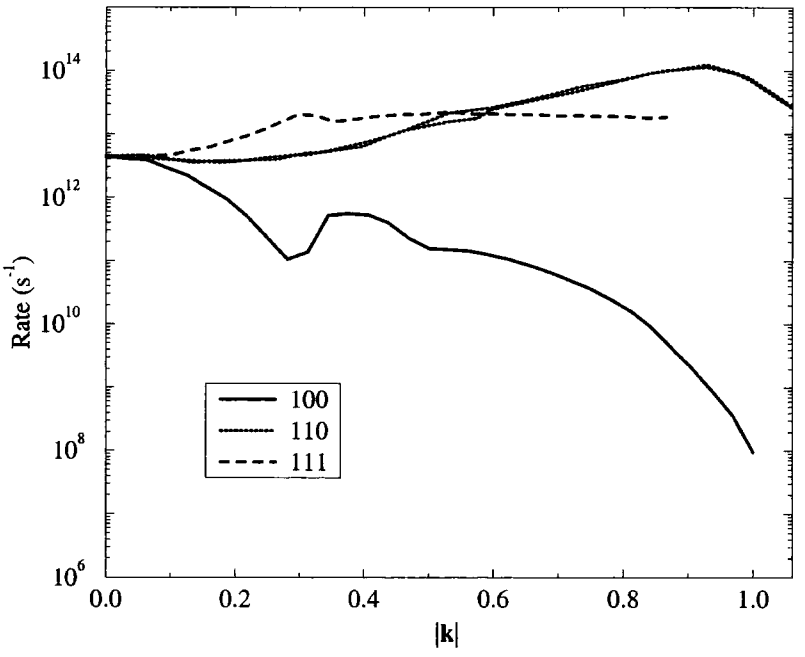


Figure 6.17: Rates in the 2nd conduction band of InGaAs plotted with respect to **k**-vector of the initiating carrier along the lines Γ–X, Γ–K and Γ–L.

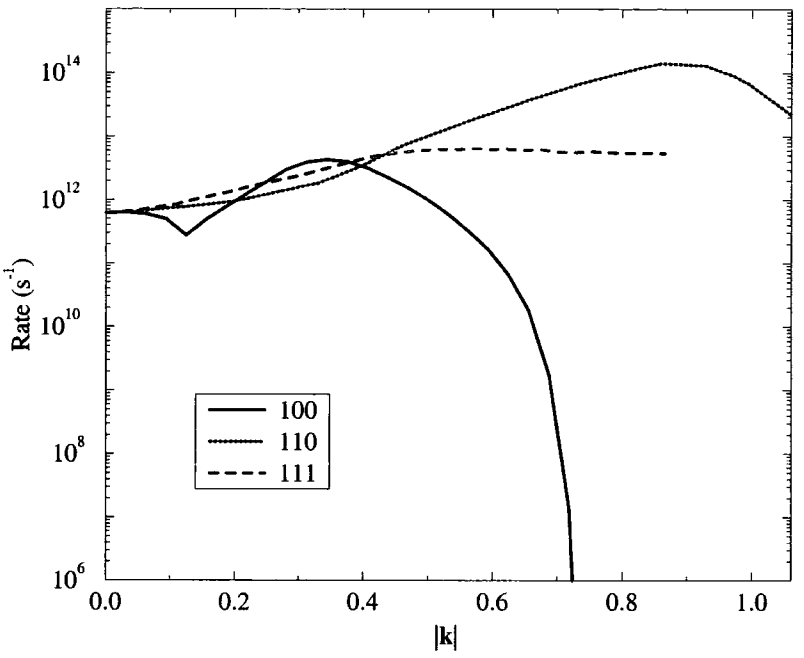


Figure 6.18: Rates in the 2nd conduction band of SiGe plotted with respect to **k**-vector of the initiating carrier along the lines Γ–X, Γ–K and Γ–L.

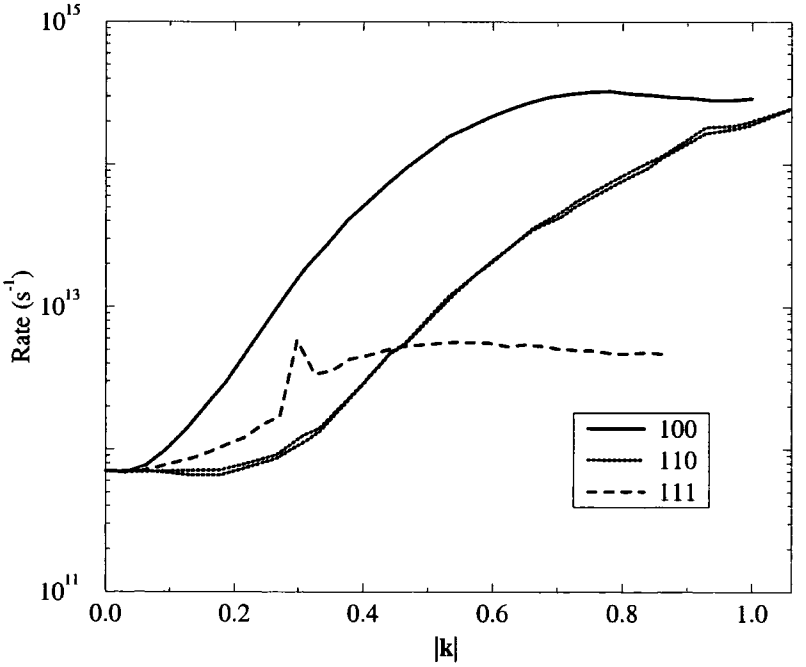


Figure 6.19: Rates in the 3rd conduction band of GaAs plotted with respect to **k**-vector of the initiating carrier along the lines Γ–X, Γ–K and Γ–L.

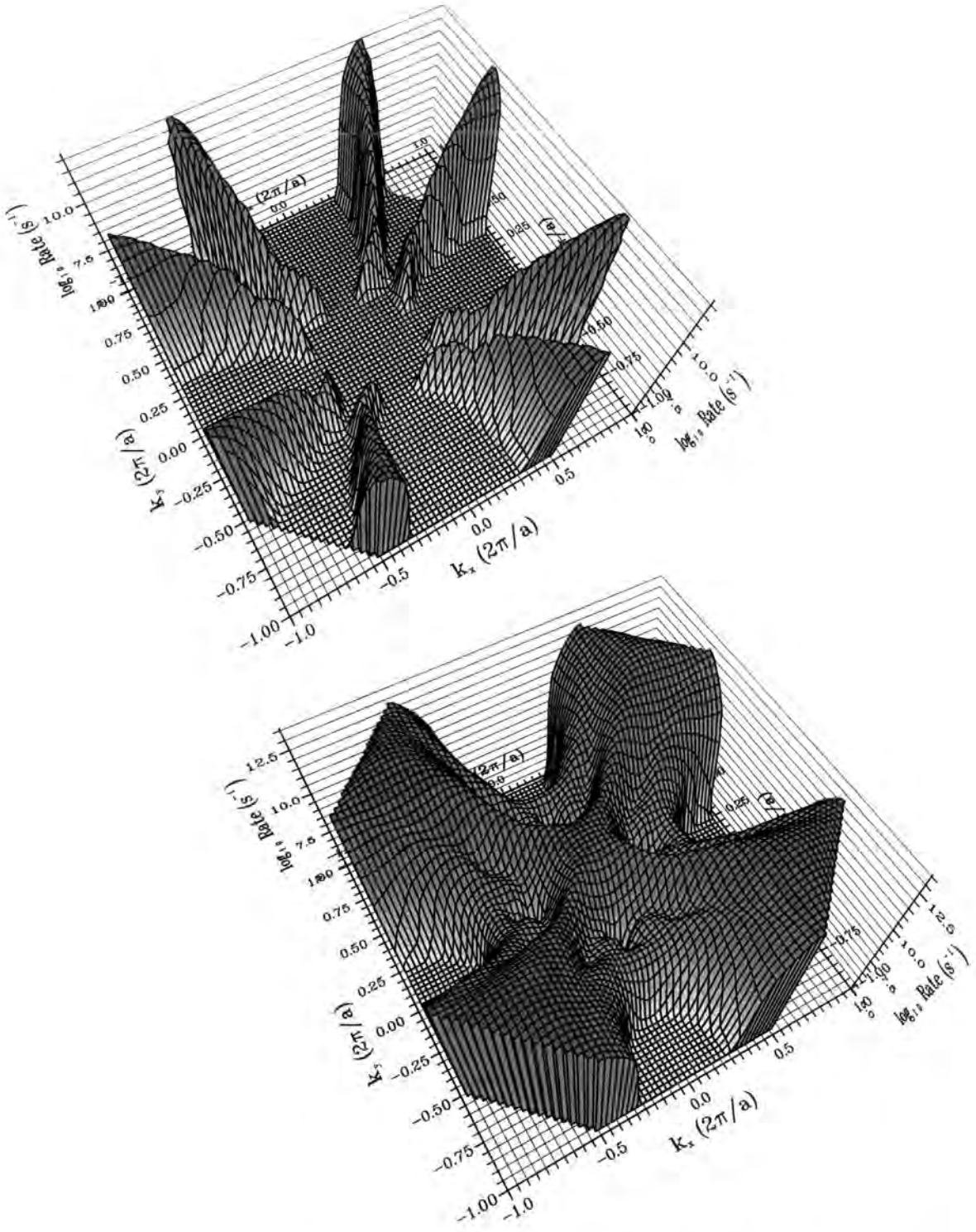


Figure 6.20: Rates in the 1st and 2nd conduction bands of GaAs plotted with respect to **k**-vector of the initiating carrier in the $k_z = 0$ plane. The rates shown here can be compared with the threshold plots in Fig. 6.7.

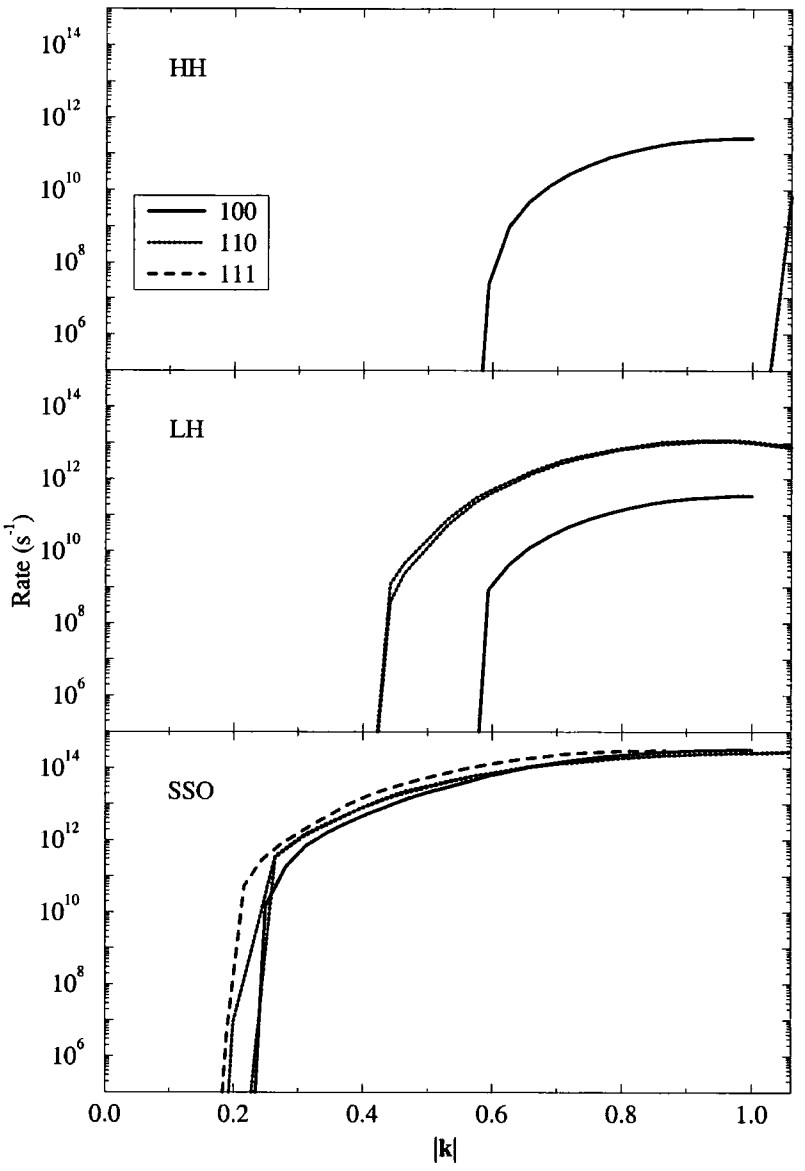


Figure 6.21: Hole initiated rates in GaAs plotted with respect to **k**-vector of the initiating carrier along the lines Γ–X, Γ–K and Γ–L.

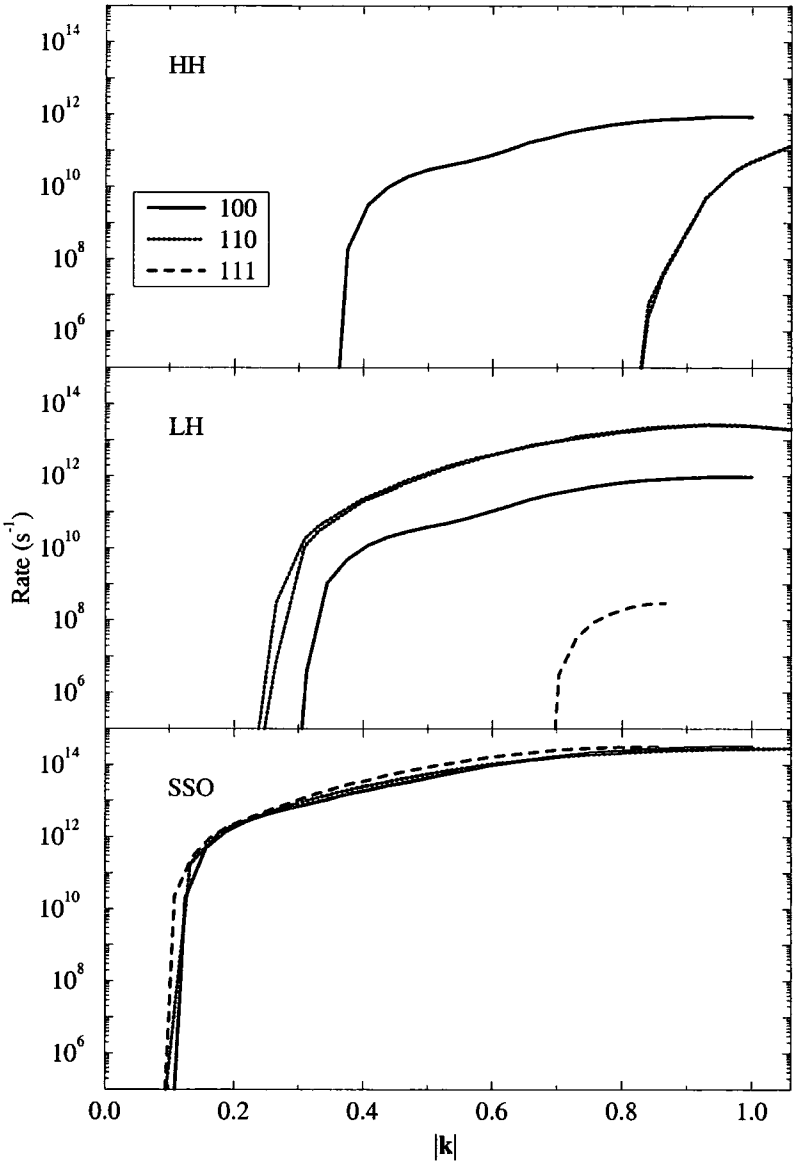


Figure 6.22: Hole initiated rates in InGaAs plotted with respect to **k**-vector of the initiating carrier along the lines Γ–X, Γ–K and Γ–L.

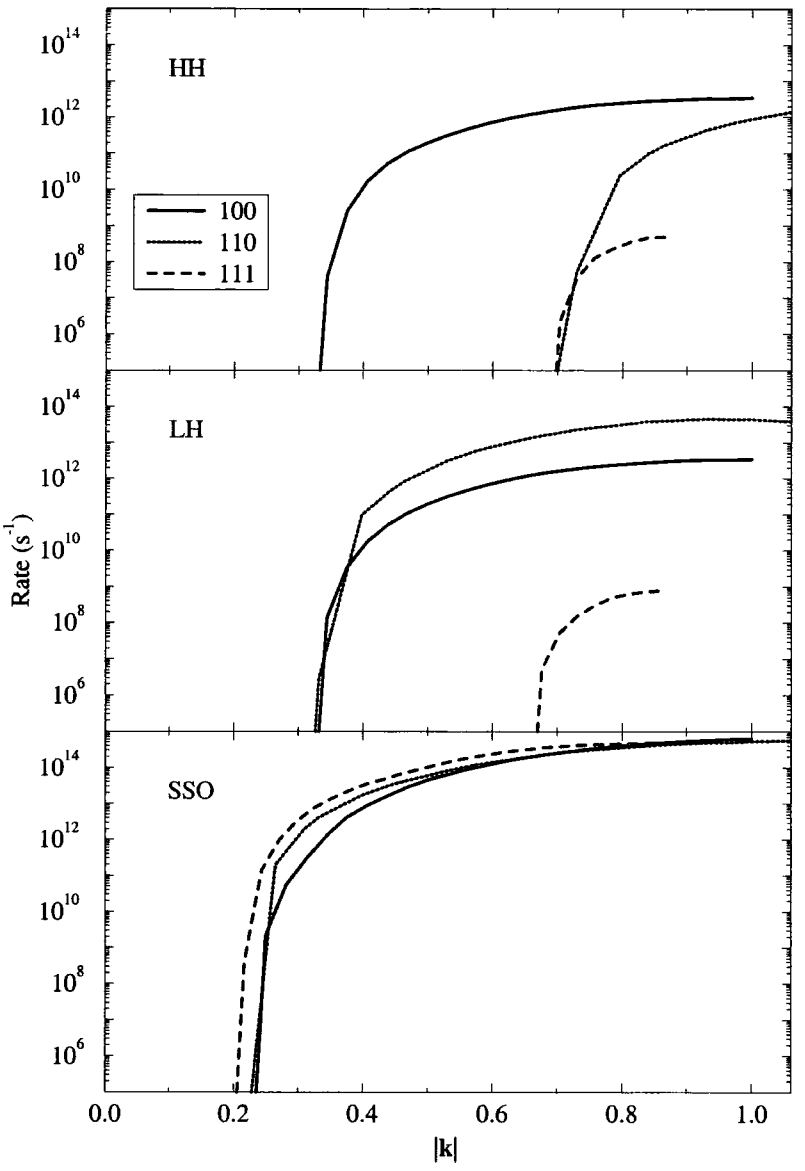


Figure 6.23: Hole initiated rates in SiGe plotted with respect to **k**-vector of the initiating carrier along the lines Γ–X, Γ–K and Γ–L.

6.4.2 Rates with respect to Energy Along Symmetry Directions

It is of particular interest to examine how the rates vary with the energy of the impacting carrier as this will be of importance in the operation of many devices. In Figs. 6.24–6.29, the same data as presented in §6.4.1 is plotted, but with the abscissa representing impacting carrier energy rather than wavevector.

In all six plots, it is clear that the rate of impact ionisation associated with a particular carrier cannot generally be expressed as a function of its energy alone — carriers at the same energy have different rates depending on their position in **k**-space.

The behaviour of the rates in each material can be compared with the behaviour of the thresholds plotted in Figs. 6.10–6.12. It was noted there that the threshold in the spin split off band showed the least anisotropy, and this is also seen in the rates which are the most nearly approximated by functions of energy alone, particularly in GaAs and InGaAs.

The threshold plot for InGaAs (Fig. 6.11) shows the $f(E)$ function having the most explicit dependence on wavevector rather than just energy in this material, particularly for the first conduction band, and this is reflected in the rates.

Another feature noticed in the threshold plots was the qualitatively different behaviour of the valence band thresholds in SiGe to that of GaAs and InGaAs. The valence band rates for these materials also differ in that their energy dependence in GaAs and InGaAs show clear differences between the bands, while in SiGe all valence bands are broadly similar.

Both electron and hole initiated rates in each material can be more accurately approximated by a function of carrier energy alone as this energy increases — a feature that will be examined in §6.4.3.

As mentioned in §6.3.1, GaAs and InGaAs have band structures of similar shape, but with InGaAs having a fundamental band gap of about half that of GaAs. Compar-

ing the plots of their electron and hole initiated rates it can be seen that for both types of carrier the rates in InGaAs have a greater spread of values at given carrier energy than in GaAs. Since the band gap of InGaAs is lower, the energy transfer in impact ionisation processes will be correspondingly lower (this is confirmed by Figs. 6.10 and 6.11 which show lower thresholds in InGaAs). The anisotropies of the bands in comparison to this will be larger in InGaAs therefore, leading to the greater variation of the rates with respect to $\mathbf{k}$ at given carrier energy.

Applying this argument to SiGe, whose band gap is closer to that of InGaAs than GaAs, would suggest that rates in SiGe should be highly $\mathbf{k}$-dependent. In fact this is not the case. Electron initiated rates in SiGe show a similar degree of $\mathbf{k}$-dependence as the wider band gap GaAs, and hole initiated rates are much more accurately expressed as a function of energy alone than hole rates in either GaAs or InGaAs. It seems likely that the cause of this difference is the indirect gap of SiGe, as compared to the direct gaps of GaAs and InGaAs. This possibility discussed again in §6.5.1.

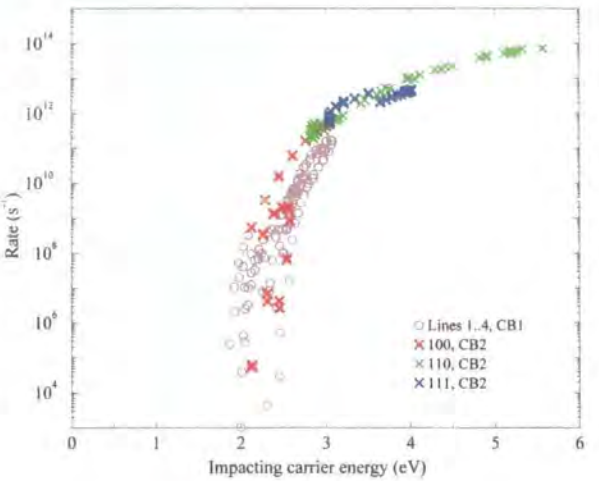


Figure 6.24: Rates of electron initiated transitions in GaAs plotted with respect to initiating carrier energy for states along the lines 1–4 in the 1st conduction band (see Fig. 6.13) and along Γ–X, Γ–K and Γ–L in the 2nd conduction band.

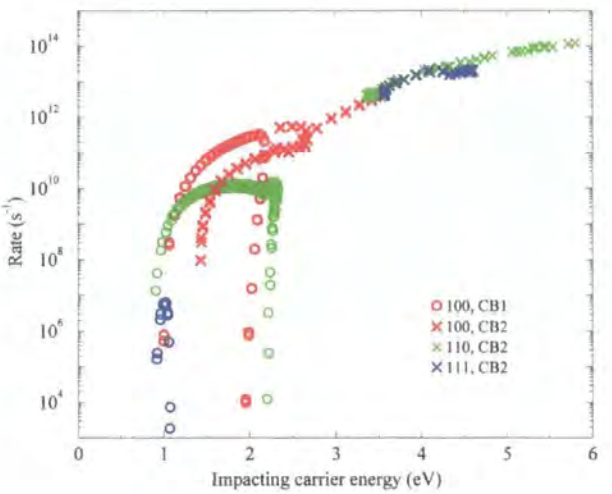


Figure 6.25: Rates of electron initiated transitions in InGaAs plotted with respect to initiating carrier energy for states along the lines Γ–X, Γ–K and Γ–L.

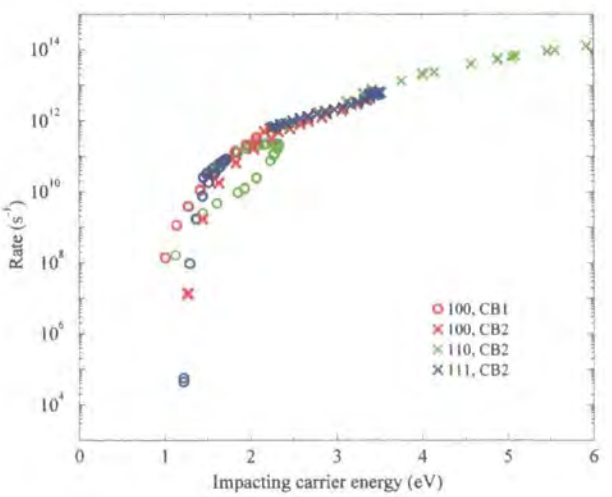


Figure 6.26: Rates of electron initiated transitions in SiGe plotted with respect to initiating carrier energy for states along the lines Γ–X, Γ–K and Γ–L.

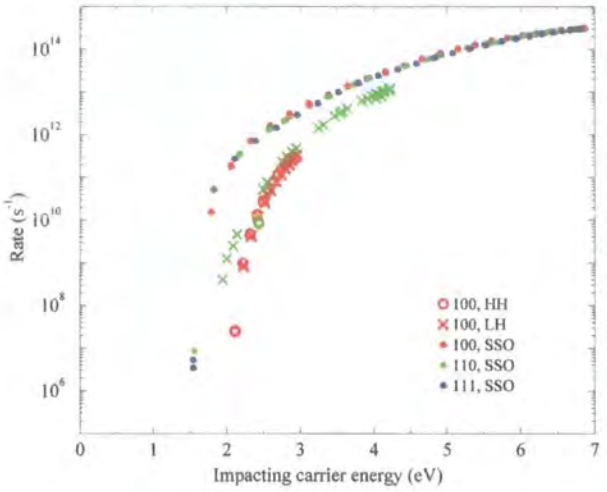


Figure 6.27: Rates of hole initiated transitions in GaAs plotted with respect to initiating carrier energy for states along the lines Γ–X, Γ–K and Γ–L.

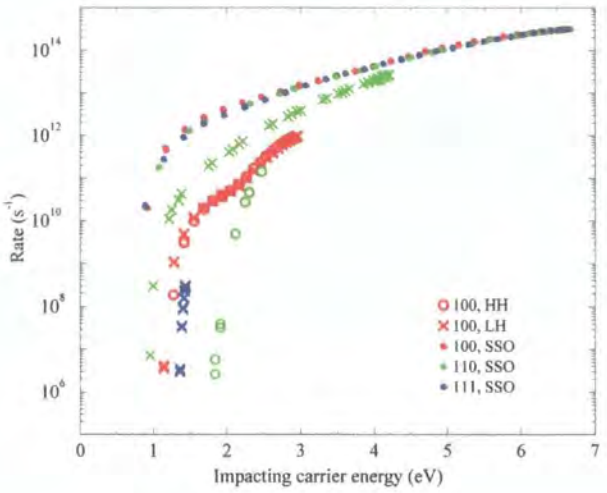


Figure 6.28: Rates of hole initiated transitions in InGaAs plotted with respect to initiating carrier energy for states along the lines Γ–X, Γ–K and Γ–L.

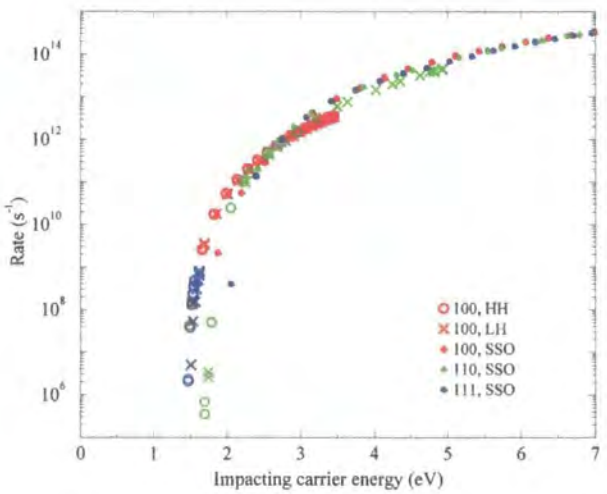


Figure 6.29: Rates of hole initiated transitions in SiGe plotted with respect to initiating carrier energy for states along the lines Γ–X, Γ–K and Γ–L.

6.4.3 Rates with respect to Energy Throughout the Zone

In §6.4.2 the rates calculated along symmetry lines of the irreducible wedge were plotted as functions of energy. To fully investigate the rate's dependence on energy, all initiating states in the Brillouin zone should be considered, and so in this section data is presented for rates calculated at **k**-vectors distributed throughout the zone[c]. Figs. 6.30–6.32 show the rates in each material plotted as a function of energy for each band. The value of the rate plotted for a given band n at a given impacting carrier energy E_i is obtained from the expression

$$R_n^{av}(E_i) = \frac{\int R_n(\mathbf{k})\ \delta(E_n(\mathbf{k}) - E_i)\ \mathrm{d}^3\mathbf{k}}{\int \delta(E_n(\mathbf{k}) - E_i)\ \mathrm{d}^3\mathbf{k}} \tag{6.3}$$

where $R(\mathbf{k})$ is the rate associated with a specific state at $\mathbf{k}$ in band n having carrier energy $E_n(\mathbf{k})$. The integrals with respect to $\mathbf{k}$ are performed over the first Brillouin zone. Thus the rate at E_i is the average rate for carriers at all **k**-vectors in band n with energy E_i.

In order to keep the required CPU time down to manageable proportions, the number of points throughout the Brillouin zone at which the rate could be calculated had to be rather limited and so some noise on the calculated results is inevitable. With more computer time, the lines presented could be smoothed out.

In Fig. 6.33 the rates for each band have been combined into average rates for electrons and holes. The average electron rate at energy E_i is obtained by taking an average of the rates in each conduction band at E_i, weighted by the corresponding density of states. A similar procedure is used for the hole initiated rates.

From the plots it can be seen that the rates for both types of carrier in all the materials are of the same order of magnitude at high energy, in agreement with observations made elsewhere [66,111]. The rates in the direct gap materials GaAs and InGaAs show similar qualitative behaviour with the spin split off band dominating at low energy. In SiGe the situation is reversed, with the rates in the conduction bands dominating at

[c]Actually, only initiating states in the irreducible wedge need be considered.

low energy. This of course corresponds to the ordering of the thresholds in these materials, as discussed in §6.3.2. (Note that the point at which the rates in Figs. 6.30–6.33 goes to zero may be less than the actual threshold value due to the finite energy width of the histogram bins used to obtain the plot).

As already noted in §6.4.2, the rate associated with a particular state in **k**-space cannot be expressed as a function of that state's energy alone. Figs. 6.34–6.36 indicate the extent to which the rates are explicitly **k**-dependent. In each figure the average rate for a band is plotted as the dark line, with the rates from individual **k**-points contributing to this average plotted as the lighter points. It can be seen that for each material, states at the same energy have rates with a range of values. The spread of individual rates generally decreases with increasing energy, also noted in §6.4.2. As was discussed in §6.4.1, at high rates the surface or surfaces of allowed transitions are large and relatively insensitive to the precise form of the band structure, and thus carriers of the same energy have similar rates. At low rates, where the energy and momentum conserving surfaces are small, the rates are highly dependent on the actual wavevector of the carrier, and so carriers at the same energy can have widely varying rates.

Electron initiated rates in InGaAs, plotted in Fig. 6.34, show the greatest spread of values at given energy. The **k**-space thresholds in the first conduction band of this material are also the most poorly defined in terms of carrier energy alone, as discussed in §6.3.2. Electron initiated rates for carriers in SiGe are comparatively well expressed in terms of the carrier energy, except near the threshold, and as noted in §6.4.2, it seems likely that this is as a result of its indirect band gap. The spread of hole initiated rates in GaAs is shown in Fig. 6.36. Rates in the spin split off band show the least explicit **k**-dependence, as is the case for the threshold in this material discussed in §6.3.2.

Although the mean rate at any given energy is generally a poor indicator of the rate due to a carrier in some specific state at that energy, particularly near the threshold, under high-field conditions carriers will be spread throughout the Brillouin zone [17,25,54]. In these circumstances, **k**-space variation in the rate will be 'integrated out' and the

overall rate due to all carriers at some particular energy will correspond reasonably well with the mean rate. Thus the mean rates plotted in Figs. 6.30–6.33 are a useful indication of the rate of ionisation in each material.

For each band of each material an analytic expression of the form

$$R(E) = A(E - E_0)^P \tag{6.4}$$

is fitted to the rate. The parameters A, P and E_0 are adjusted to give the best fit as follows. Taking the logarithm of both sides of Eq. (6.4) gives a straight line of the form $y = ax + b$ where $y = \log R$, $x = \log(E - E_0)$, $a = P$ and $b = \log A$. The values of a and b giving the best fit by least squares to $y(x)$ are determined for a fixed value of E_0. The fit has an associated RMS error, which can itself be minimised by adjusting E_0. Table 6.7 lists the fitting parameters obtained for the rates in each band of each material. Note that the small spin splitting of each band has been neglected for these parameters, which are fitted to the average rate for each band pair. Note also that the fitted value of E_0 is obtained without reference to the actual threshold energies, which are also listed in Table 6.7.

To put the parameters listed in Table 6.7 in context, consider the fits obtained from idealised band structure consisting of spherical parabolic bands, and for which the matrix elements are constant. For the case of a direct gap Keldysh [53] calculated that $P = 2$, and for an indirect gap Beattie obtained the value $P = 3$ near threshold. In these cases, the value of A then determines how hard or soft the threshold is. The impact ionisation threshold is described as hard if carriers ionise very quickly once the threshold energy has been achieved, and conversely a soft threshold corresponds to the case in which carriers are not immediately ionised upon reaching the threshold but can survive to considerably higher energies. Thus, the higher the value of A, the greater is the ionisation rate above threshold and so the harder the threshold is.

Examining the parameters presented in Table 6.7, it is clear that for both carrier types in all materials, the values of P obtained are significantly higher than those

obtained from the calculations based on the idealised band structure, particularly in the case of the direct gap materials. Rate calculations based on full band structure typically obtain such higher P-values [26,28,67]. A higher P-value, as well as indicating greater deviation of the real band structure from the idealised case, is also an indication of a softer threshold [22,66]. Thus the use of real band structure significantly increases threshold softness over that obtained from the Keldysh formula. Values of A listed in the table range over about an order of magnitude, with the largest being for electrons in GaAs indicating the hardest threshold. However, when P-parameters differ, the A-parameters are not strictly comparable. Thus threshold softness is influenced by the combination of A and P values.

Note that the characteristics of the rate alone do not determine whether the threshold is hard or soft, but only act as a guide to what we might expect. The question of whether carriers ionise quickly upon reaching threshold or continue to higher energies can only be answered by considering in detail the transport of carriers in the material, including the effects of the real band structure and other scattering mechanisms, particularly phonon scattering. Monte Carlo simulation is an appropriate technique to perform these calculations.

Material	Band	A	P	E_0 (fit)	E_0 (calc)
	SSO	1.4×10^{12}	3.2	1.51	1.51
	LH	1.4×10^{11}	4.6	1.71	1.75
	HH	2.8×10^{11}	4.4	1.99	1.98
GaAs	CB 1	8.5×10^{09}	8.7	1.68	1.85
	CB 2	2.2×10^{11}	4.7	1.91	2.00
	e^-	1.4×10^{11}	5.2	1.89	1.85
	h^+	8.2×10^{10}	5.1	1.43	1.51
	SSO	2.4×10^{12}	2.6	0.75	0.78
	LH	1.0×10^{11}	4.4	0.84	0.93
	HH	2.6×10^{10}	5.4	1.03	1.20
InGaAs	CB 1	1.3×10^{10}	5.6	0.75	0.87
	CB 2	1.3×10^{11}	4.3	1.07	1.43
	e^-	1.6×10^{10}	5.6	0.75	0.87
	h^+	1.5×10^{11}	4.2	0.73	0.78
	SSO	8.2×10^{11}	3.5	1.71	1.65
	LH	2.3×10^{11}	4.1	1.39	1.41
	HH	7.3×10^{10}	5.2	1.22	1.27
SiGe	CB 1	2.7×10^{10}	5.1	0.81	0.91
	CB 2	1.5×10^{11}	4.1	0.95	0.95
	e^-	4.6×10^{10}	4.9	0.84	0.91
	h^+	7.8×10^{10}	4.7	1.23	1.27

Table 6.7: Fitting parameters for the rates shown in Figs. 6.30–6.33. The fitted value of the rate is given by: $R(E) = A(E-E_0)^P$ (with R in units of s^{-1} and E in eV).

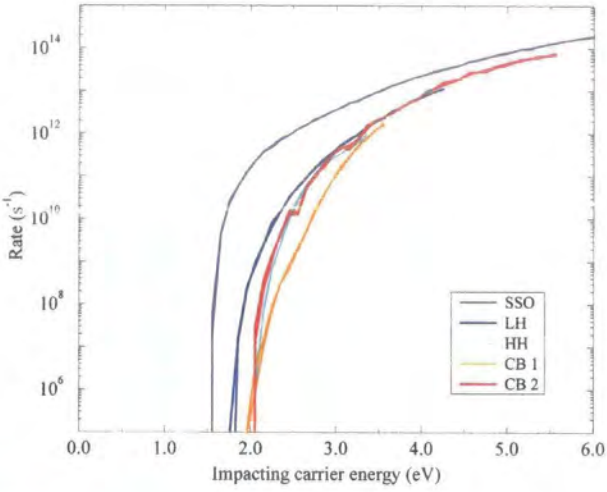


Figure 6.30: Rates in GaAs, averaged throughout the Brillouin zone.

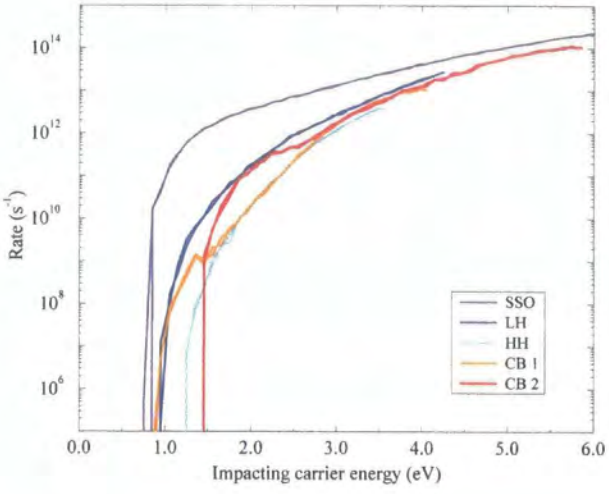


Figure 6.31: Rates in InGaAs, averaged throughout the Brillouin zone.

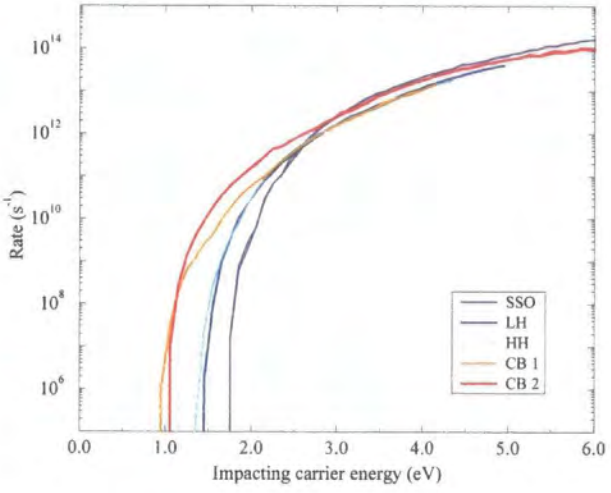


Figure 6.32: Rates in SiGe, averaged throughout the Brillouin zone.

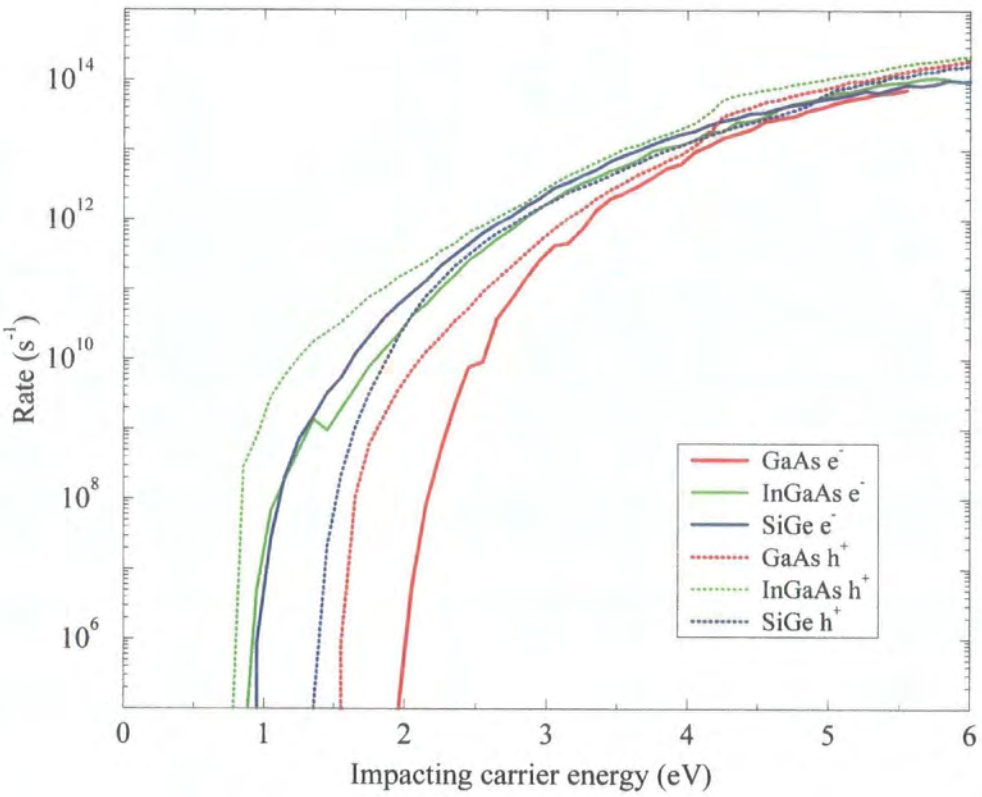


Figure 6.33: Rates for electrons and holes in each material, averaged throughout the Brillouin zone.

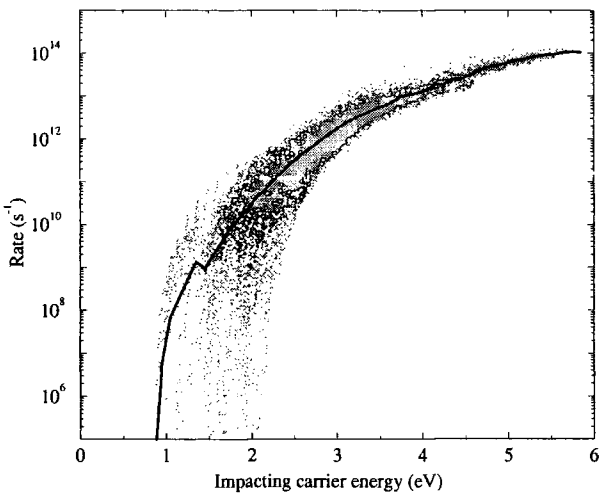


Figure 6.34: Spread of electron initiated rates in InGaAs. The dark line is the averaged rate as a function of energy. The lighter points are the individual rates evaluated at specific **k**-vectors.

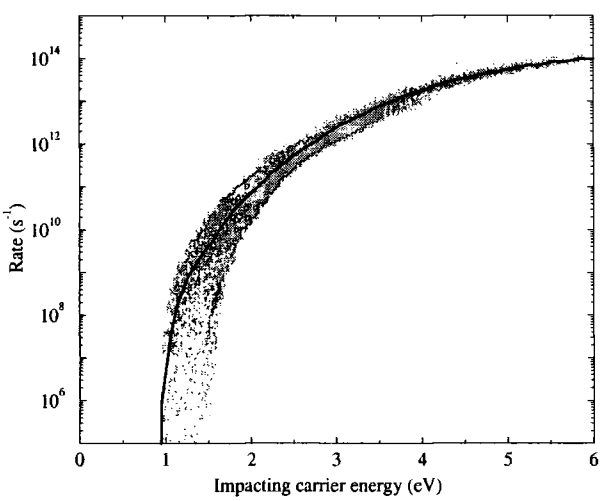


Figure 6.35: Spread of electron initiated rates in SiGe.

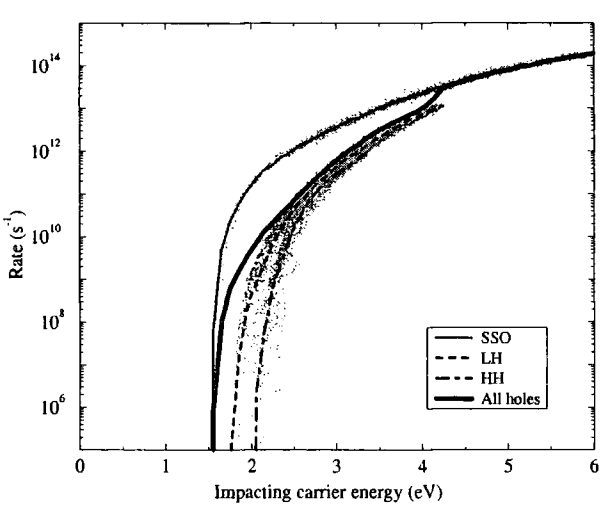


Figure 6.36: Spread of hole initiated rates in GaAs.

6.5 Generated Carriers

The distribution of the carriers generated by the impact ionisation process is of interest, both in understanding the factors influencing the rate and from the point of view of transport simulations. In §6.1, the distinction between *secondary states* and *generated carriers* was noted, in particular that the wavevector of the impacted state is minus that of the corresponding generated carrier. To avoid confusion, the discussion in §6.5.1 below is limited strictly to the distributions of secondary states only. The corresponding distributions of generated carriers can be straightforwardly obtained from these according to the considerations outlined in §6.1.

Since time reversal symmetry ensures that $E(\mathbf{k}) = E(-\mathbf{k})$, §6.5.2 and §6.5.3 which discuss the energies of the generated carriers do not need to be concerned with the distinction between secondary states and generated carriers.

6.5.1 Distribution in k-Space of Secondary States

Figs. 6.37–6.42 are all of the same type. Each figure consists of a line graph at the top of the page and two rows of five surface plots below it. The line graph is a rate plotted with respect to the wavevector of the initiating state along a symmetry line in the Brillouin zone. The base of each of the ten octagonal plots is the $k_z = 0$ plane of the Brillouin zone with the height of the plot indicating the density of secondary states in **k**-space, projected onto this plane[d]. The density of secondary states is obtained by considering all states sampled in the Monte Carlo rate integration, weighted by the corresponding matrix elements. Spaced along the rate graph are five circles indicating the rates due to specific initiating **k**-states. The upper row of surface plots shows the final states corresponding to these five circles and the lower row shows the impacted states. The left-most upper and lower surface plots correspond to the left most circle, and so on from left to right.

[d]Due to projecting secondary state density onto the $k_z = 0$ plane, states lying in the valleys at 001 and $00\bar{1}$ appear to lie at Γ. On a 2-dimensional plot this ambiguity unfortunately cannot be avoided.

Figs. 6.37 and 6.38 are plotted along the Γ–X and Γ–K lines respectively for electron initiated transitions in InGaAs of the type CB2,HH→CB1,CB1 (as defined in §5.5 of Chapter 5) i.e. the final states lie in the first conduction band and the impacted states lie in the heavy hole band. For the initiating states along Γ–X, where the rates are lower, the distribution of final states is sharply peaked in the valley bottoms. The impacted states also lie at the position of lowest carrier energy, i.e. near Γ. For impacting states along Γ–K, where the rates are generally higher, the final states are located throughout the Brillouin zone, although preferentially in the conduction band valleys at all but the highest rates. The impacted states are similarly located throughout the zone — the roughly square shape to the distribution is as a result of the states in the L-directions corresponding to low energy holes being favoured.

Figs. 6.39 and 6.40 show the positions of secondary states for CB1,HH→CB1,CB1 type transitions in SiGe, with impacting carriers located along the Γ–X and Γ–K lines respectively. Since the Γ-valley of SiGe is very shallow, in each case no final states lie there, being located generally in the X-valleys (recall that the projection of states onto the $k_z = 0$ plane means states in the 001 and $00\bar{1}$ valleys appear to lie at Γ). As with InGaAs, the impacted states tend to lie towards Γ where hole energy is lowest, and also have the square-shaped distribution which is as a result of the low energy L-directions being favoured.

In §6.4.3, Figs. 6.34 and 6.35 were compared and it was noted that the electron initiated rates in SiGe could be much better fitted by a function of energy alone than could the electron rates in InGaAs. The distributions of secondary states plotted in Figs. 6.37 and 6.38 for InGaAs and Figs. 6.39 and 6.40 for SiGe hint at the cause of this better fit in SiGe. Comparing the plots for the X- and K-directions in InGaAs, it can be seen that the distributions of secondary states differ considerably between the two. Along Γ–X both the impacted and final states are located at sharp peaks near the minima of their respective bands. In the Γ–K direction however, secondary states corresponding to both types of carrier are distributed widely throughout the zone. A

similar comparison of plots for the X- and K-directions in SiGe reveals much greater similarity in the distributions of secondary states, despite the range of magnitudes of rates involved being no smaller than for InGaAs. The X-valleys of the first conduction band are well populated by the final states in both directions at all rates but the very lowest. Similarly the impacted states are not sharply peaked as along the X-direction of InGaAs, but are distributed about Γ in a similar way for both lines.

For low energy impacting states in InGaAs, it seems likely that the highly localised distributions of impacted and final states leads to rates that are sensitive to the initiating state's specific **k**-vector. In SiGe, the secondary states are distributed throughout all the X-valleys even at low energy, and the qualitative form of this distribution does not change greatly as the impacting state energy increases. Thus the rates in SiGe are less sensitive to the initiating carrier's specific **k**-vector.

In Fig. 6.41, secondary states involved in hole initiated impact ionisations in GaAs are plotted. Impacting vectors lie in the spin split off band along the Γ–X line, and make transitions of the type SSO,CB1→HH,HH, i.e. the final states lie in the heavy hole band and generated electrons lie in the first conduction band. As in the case of electron initiated transitions, final states lie at low energy (i.e. near Γ) at low rate, and tend to lie throughout the heavy hole band as the rate increases. Again, the secondary states lying in the valence band have a roughly square distribution when projected onto the $k_z = 0$ plane due to favouring the low energy L-directions in the heavy hole band. The impacted state distribution is sharply peaked in the Γ-valley at the lowest rates. As the rate increases, the distribution spreads but remains peaked in the conduction band valleys. Only at the highest rates do the impacted states lie throughout the Brillouin zone.

Fig. 6.42 shows the same hole initiated data as in Fig. 6.41, but plotted for SiGe instead of GaAs. The final state distribution is qualitatively very similar to that of GaAs. At low rate the impacted states are located near the minima of the conduction band as is also the case in GaAs. However, in SiGe this means that near threshold

impacted states are located at X rather than at Γ as in GaAs. As was discussed in §6.3.2, the fact that at threshold, impacted states do not lie at Γ leads to the qualitatively different behaviour of the thresholds in SiGe compared to the direct gap materials GaAs and InGaAs.

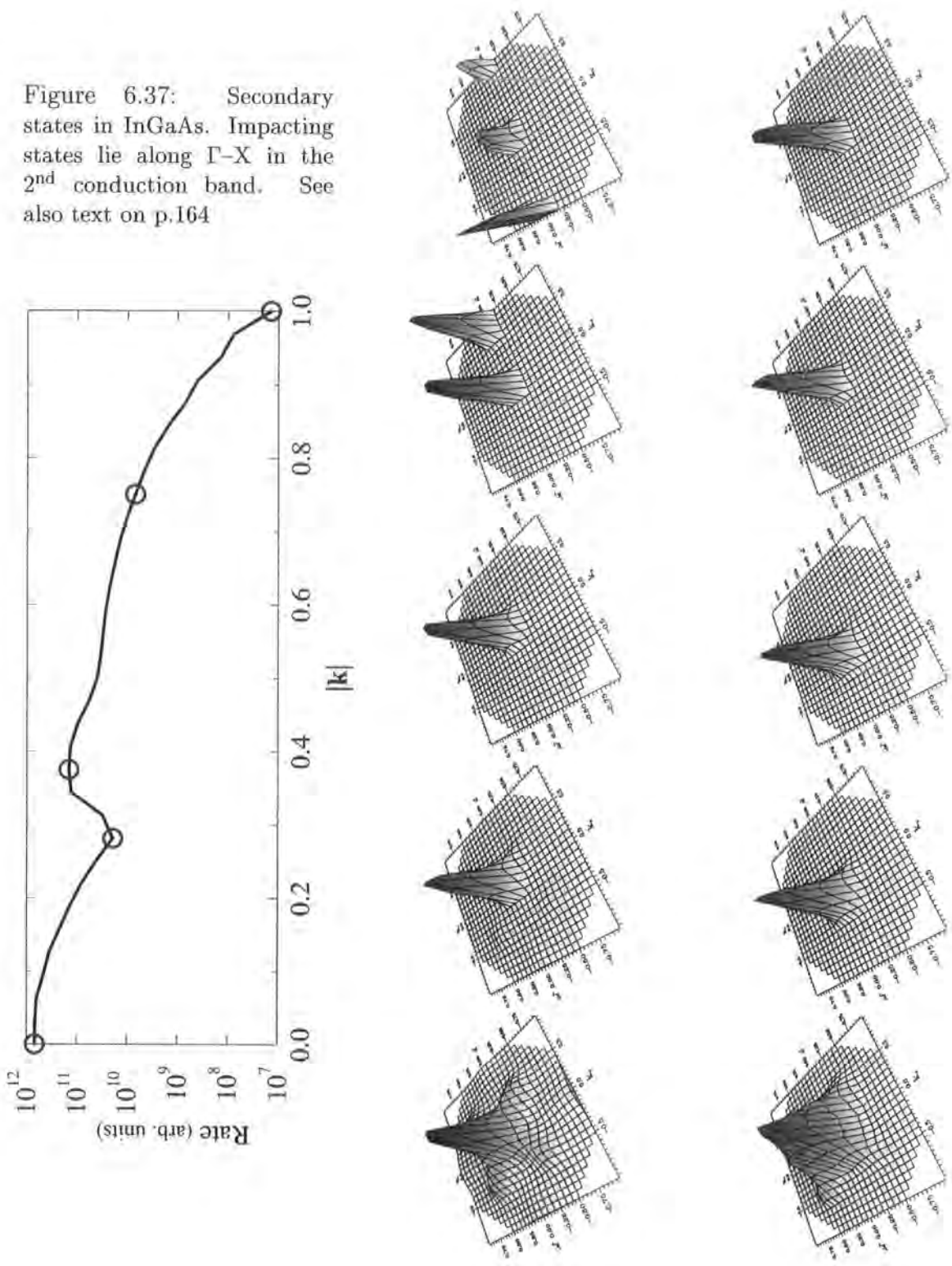


Figure 6.37: Secondary states in InGaAs. Impacting states lie along Γ–X in the 2nd conduction band. See also text on p.164

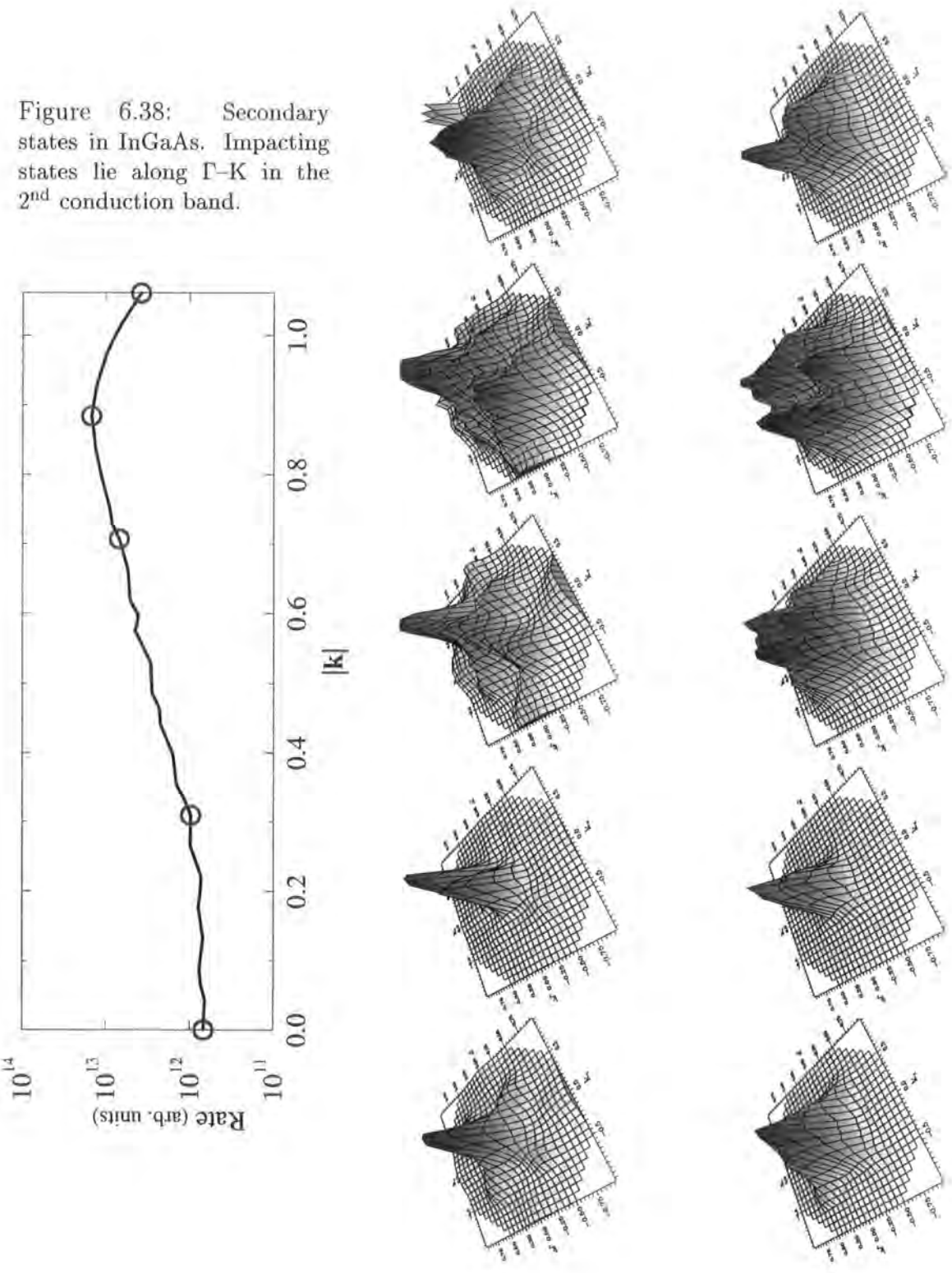
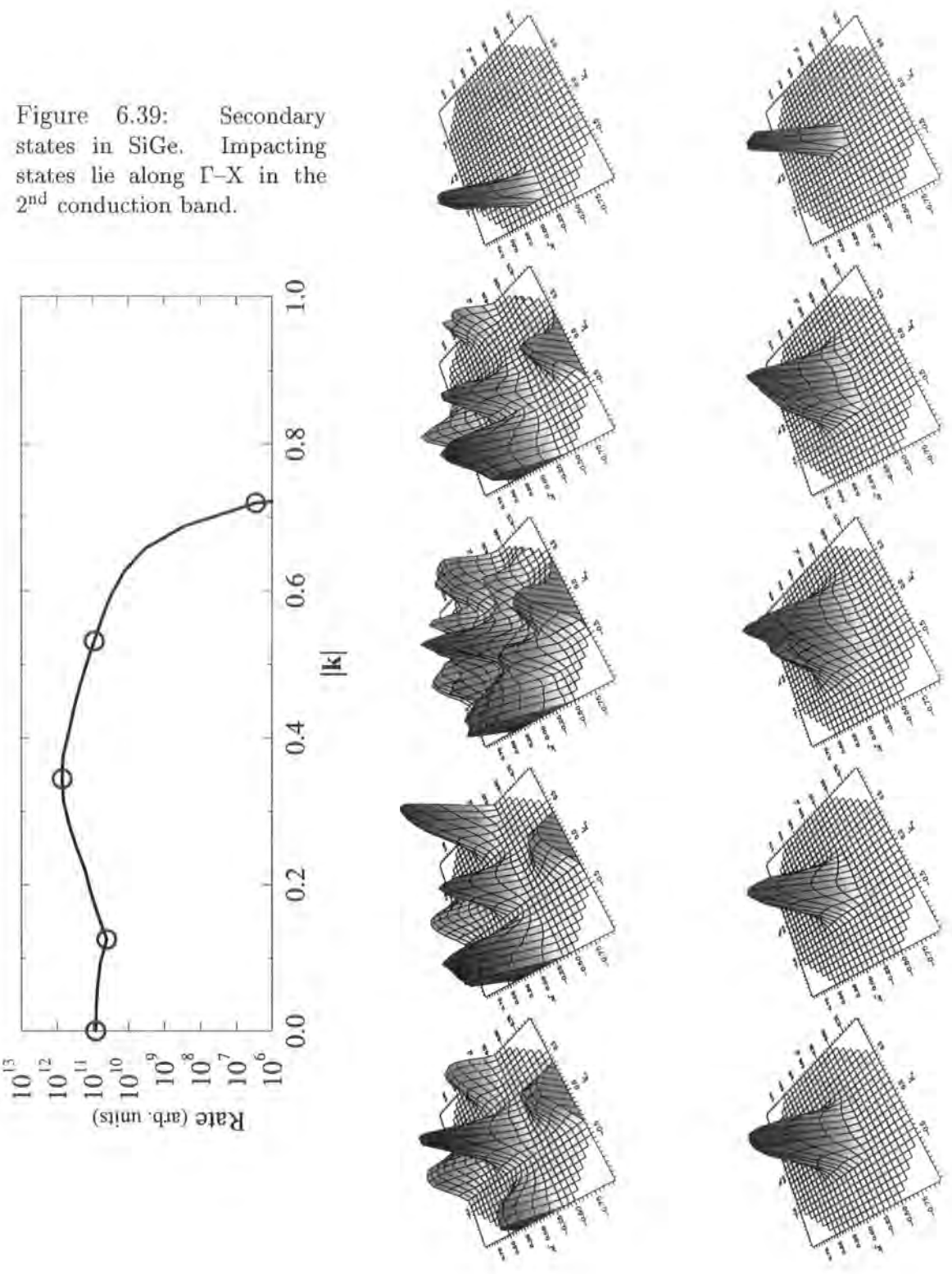


Figure 6.38: Secondary states in InGaAs. Impacting states lie along Γ–K in the 2nd conduction band.

Figure 6.39: Secondary states in SiGe. Impacting states lie along Γ–X in the 2nd conduction band.



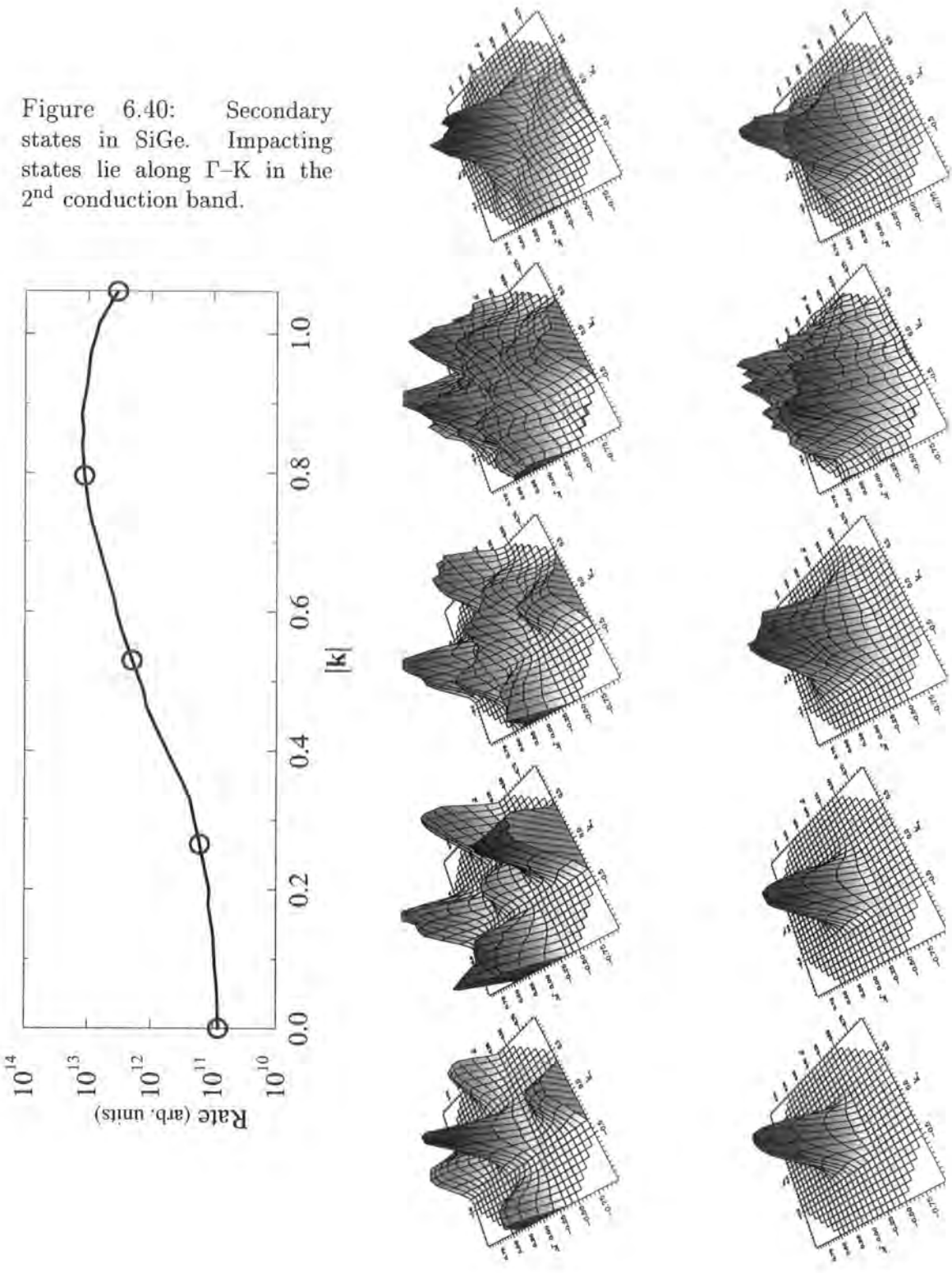


Figure 6.40: Secondary states in SiGe. Impacting states lie along Γ–K in the 2^{nd} conduction band.

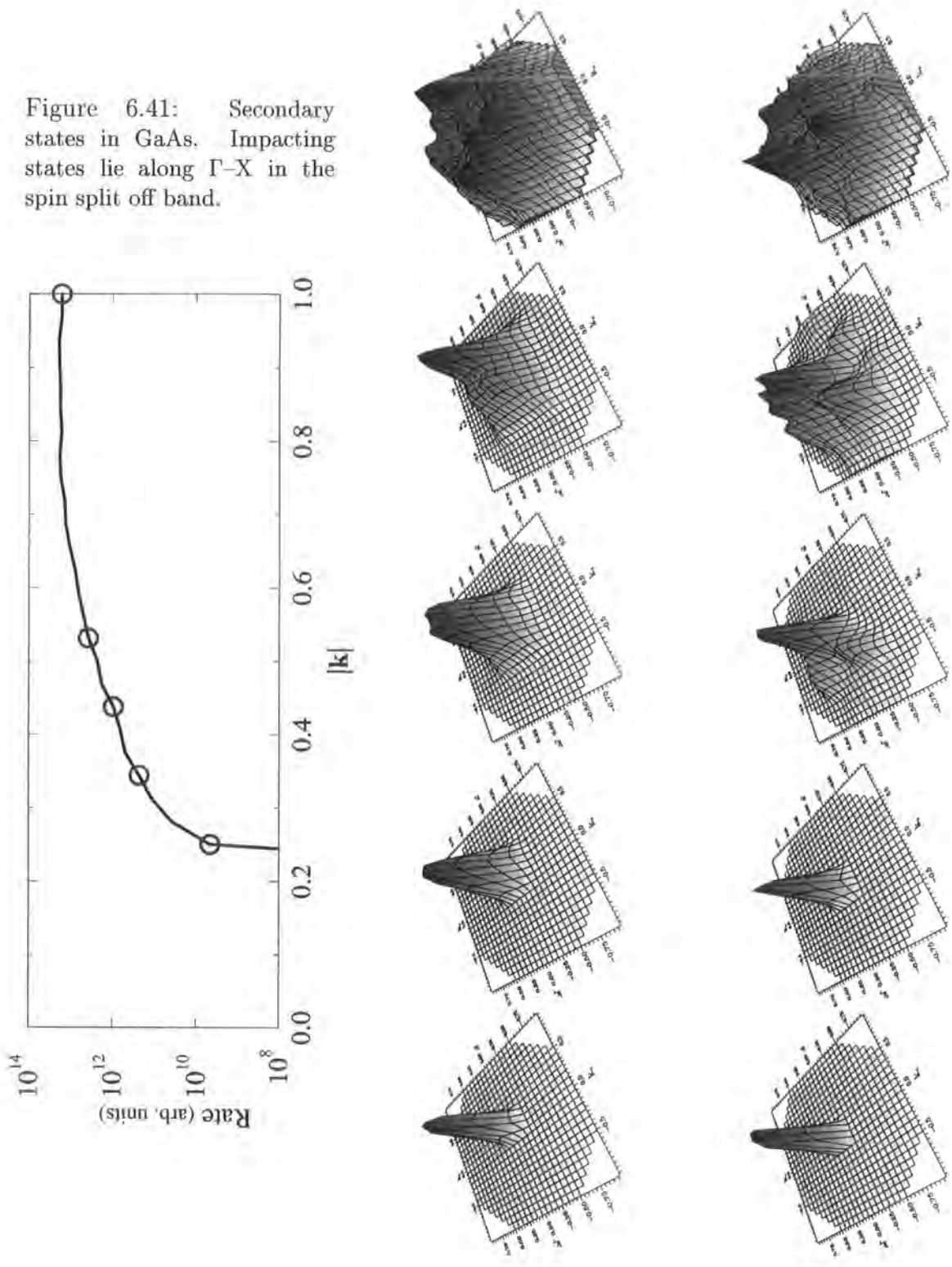


Figure 6.41: Secondary states in GaAs. Impacting states lie along Γ–X in the spin split off band.

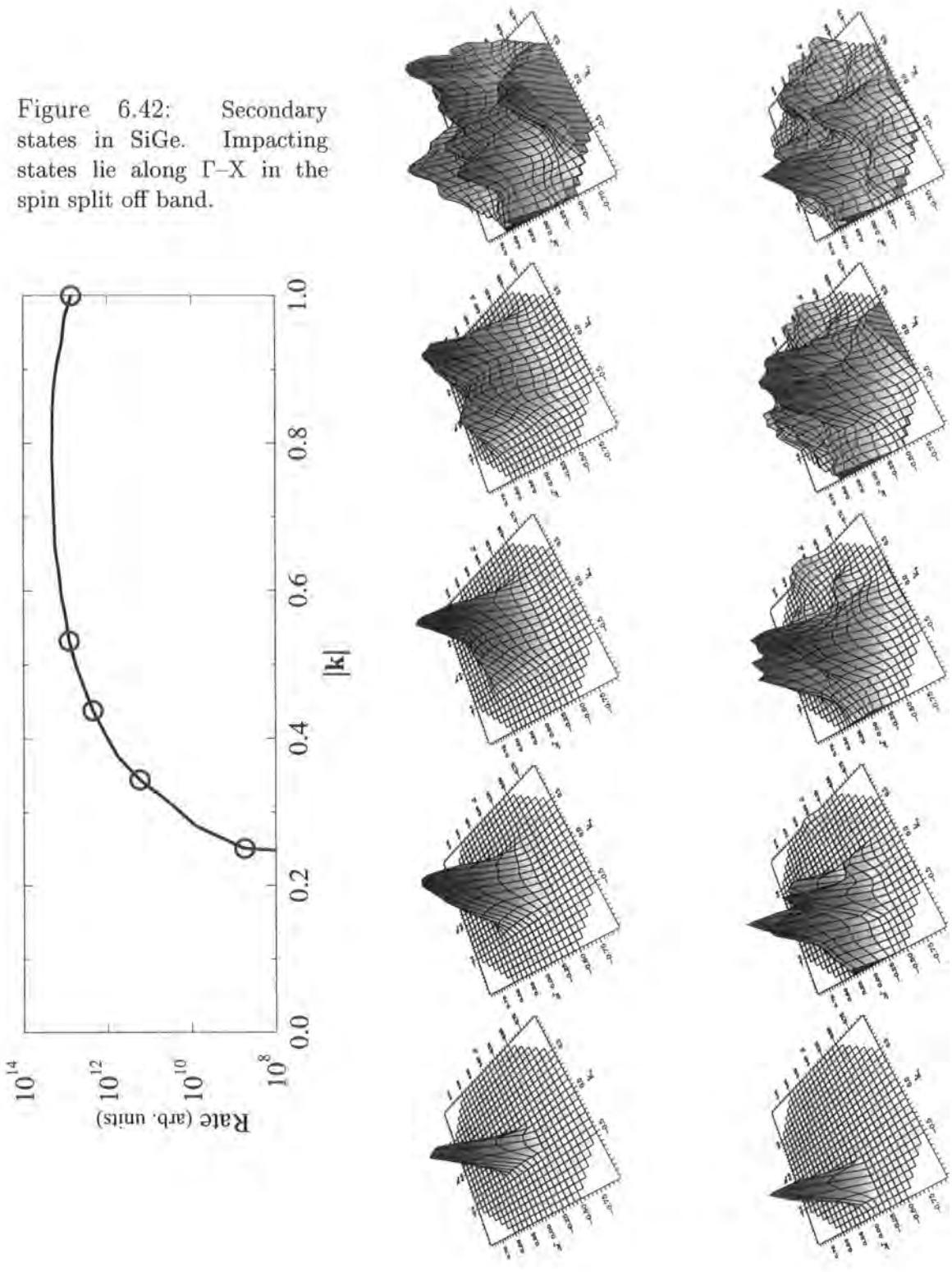


Figure 6.42: Secondary states in SiGe. Impacting states lie along Γ–X in the spin split off band.

6.5.2 Mean Energies of Generated Carriers

The mean energies of the generated carriers are presented here as a function of the initiating carrier energy. The mean energy of the secondary final states E_{sf} is calculated for an impacting carrier of energy E_i using the expression

$$\overline{E_{sf}}(E_i) = \frac{\sum_n \frac{1}{2}(E_{n,\mathbf{k}_{1'}} + E_{n,\mathbf{k}_{2'}})|M_n|^2}{\sum_n |M_n|^2} \tag{6.5}$$

where the sum is over all pairs of final states sampled in the Monte Carlo integration, $E_{n,\mathbf{k}_{1'}}$ and $E_{n,\mathbf{k}_{2'}}$ are the energies of the n^{th} sampled pair and $|M_n|^2$ is the squared modulus of the corresponding matrix element for the transition. Thus in calculating the mean energy of the final states, each pair is weighted by the corresponding matrix element, reflecting the probability that if a transition occurs, it will be made to that specific pair of states. Note that the Monte Carlo integration procedure itself accounts for the effect of the density of final states. The mean energy of the impacted secondary states is calculated similarly using the expression

$$\overline{E_{si}}(E_i) = \frac{\sum_n E_{n,\mathbf{k}_2}|M_n|^2}{\sum_n |M_n|^2} \tag{6.6}$$

Figs. 6.43 and 6.44 show the mean energies of the carriers generated by transitions from states in the first and second conduction bands of InGaAs. The colour of the points on the plots indicates which symmetry line the initiating carrier is located on. In the second conduction band plot, an approximately linear relation between the energies of the impacting carrier and generated carriers can be seen. In the first conduction band, i.e. at lower energy, the relation is much less clear; as with rates, the generated carrier energy is more sensitive to the actual **k**-vector of the initiating carrier at low energy. The difference in energies plotted along the 100- and 110-directions in the first conduction band is of the order of several percent. This should be compared with the two orders of magnitude separating the corresponding rates, plotted in Fig. 6.25. The deviation of the energies of the generated holes from a simple function of impacting carrier energy is greater than that for generated electrons. This is to be expected, since

energy conservation requires that

$$2\overline{E_{sf}} + \overline{E_{si}} = E_i - E_g \tag{6.7}$$

where E_g is the energy gap. Thus a deviation in the mean secondary final state energy $\overline{E_{sf}}$ must be accounted by a deviation in the mean secondary impacted state energy $\overline{E_{si}}$ of twice the magnitude.

Figs. 6.45–6.47 compare the mean energies of generated carriers for electron and hole initiated impact ionisation in each material. As with other aspects of impact ionisation examined, the behaviour of GaAs and InGaAs is qualitatively similar, while that of SiGe differs. In the direct gap materials, the generated electrons in both electron and hole initiated transitions tend to take the slightly greater share of the available energy. In contrast, for low impacting energy in SiGe, the energies of impacted and final states are similar, while at higher impacting energies the generated holes tend to be of slightly greater energy. In SiGe, generated carrier energies are more accurately represented by a function of energy only, while in GaAs and InGaAs, they are more explicitly **k**-dependent, as is also the case with the rates themselves.

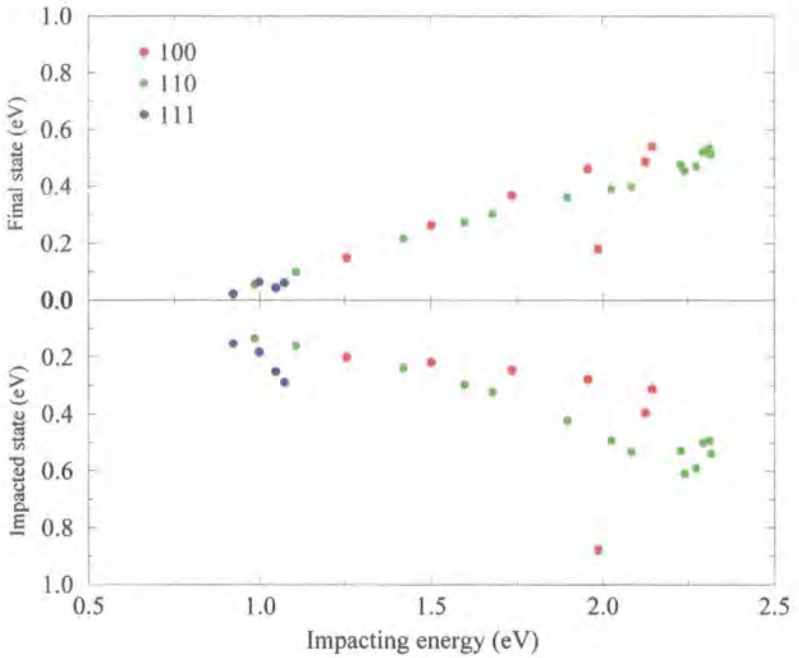


Figure 6.43: The mean energy of carriers generated by impacting electrons located along the 100, 110 and 111 directions in the 1st conduction band of InGaAs

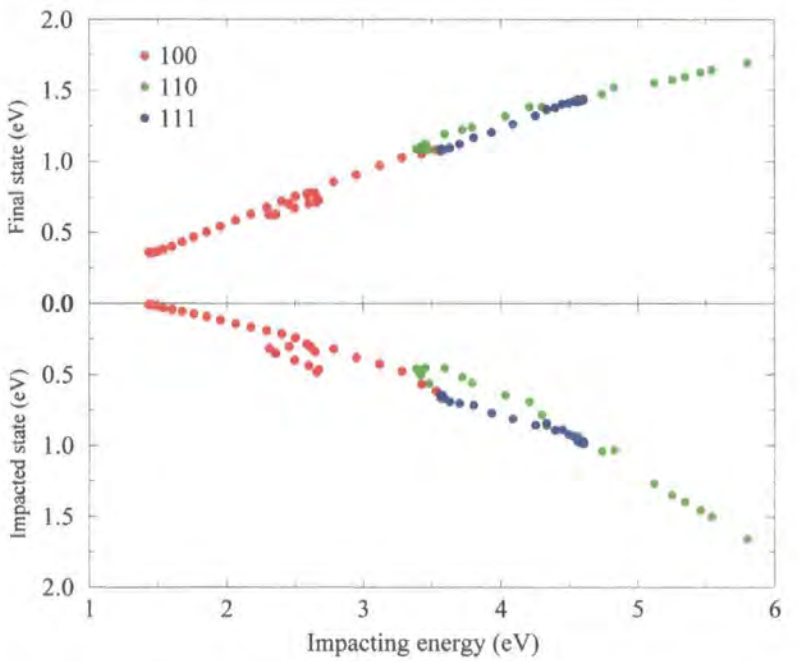


Figure 6.44: The mean energy of carriers generated by impacting electrons in the 2nd conduction band of InGaAs.

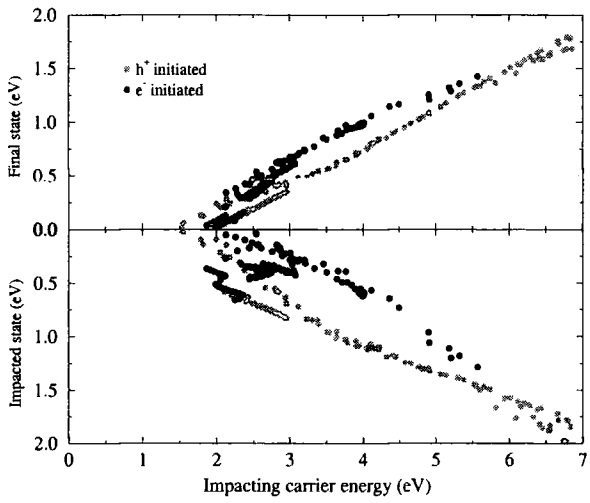


Figure 6.45: The mean energies of generated carriers for electron and hole initiated impact ionisation in GaAs.

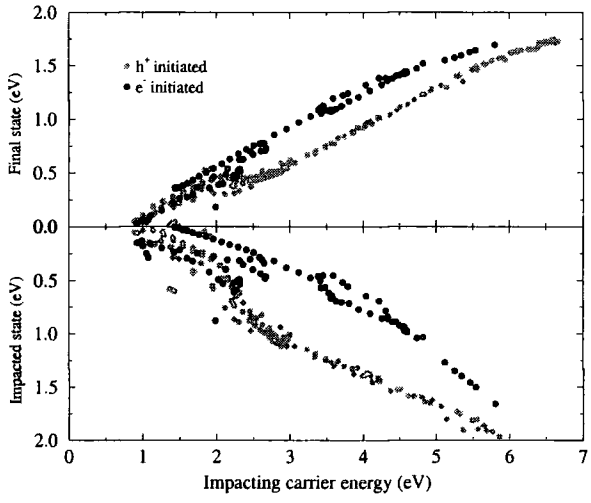


Figure 6.46: The mean energies of generated carriers for electron and hole initiated impact ionisation in InGaAs.

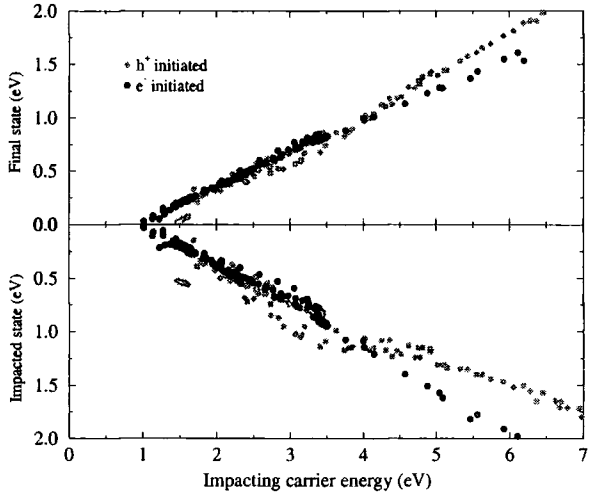


Figure 6.47: The mean energies of generated carriers for electron and hole initiated impact ionisation in SiGe.

6.5.3 Distribution of Energies of Generated Carriers

In §6.5.2 the mean energies of the generated carriers were presented. Naturally, a given impacting carrier does not generate secondary carriers with a specific energy but rather with a distribution of energies. The form of the distribution is examined here.

Figs. 6.48–6.53 show the energy-distribution of generated carriers as a function of the impacting carrier energy. The height $f(E_i, E_s)$ of each plot shows in arbitrary units the number of secondary carriers generated with energy E_s by an impacting carrier of energy E_i. The distribution of generated carriers is calculated using the final state pairs considered in the Monte Carlo integration of the rate, each weighted by the corresponding matrix element. The plots are normalised so that

$$\int_0^\infty f(E_i, E_s)\, \mathrm{d}E_s \quad = \quad C \tag{6.8}$$

where C is an arbitrary constant.

In Figs. 6.48 and 6.49, the distributions of final and impacted states are plotted for impacting electrons in the second conduction band of SiGe. The distributions of each type of secondary carrier are similar, with the generated hole distribution being at slightly higher energy for the more energetic impacting carriers. This confirms the observations made for the mean energies of these carriers, plotted as the dark circles in Fig. 6.47.

Figs. 6.50 and 6.51 show the corresponding plots for impacting electrons in the second conduction band of InGaAs. At low energy the distribution of final states (Fig. 6.50) has a doubly peaked structure: the lower energy peak corresponds to final states in the Γ-valley, while the higher energy peak corresponds to the X-valley. Remembering that at low energy, impacting carriers in the second conduction band lie along the Γ–X line, this observation is confirmed by Fig. 6.37 which shows the **k**-space distribution of secondary states for these carriers. At higher impacting electron energies, the structure of the distribution is smoothed out to a single flatter peak. The distribution of generated holes (Fig. 6.51) shows none of the complexity of the gener-

ated electrons as a result of the valence band structure being correspondingly simpler. As was noted for the plots of mean carrier energy (Fig. 6.46), the electrons are generated with generally slightly higher energies than the holes. Note that although the bottom of the X-valley lies 0.67 eV above the conduction band edge (see Table 6.4), final states lying within the upper peak of Fig. 6.50 appear to have energies down to about 0.55 eV. This is as a result of the finite bin width of the histogram used to calculate the carrier distribution, and the interpolation algorithm applied by the plotting package.

Figs. 6.52 and 6.53 show final and impacted states for impacting carriers in the spin split off band of GaAs. The distribution of generated electrons corresponding to the impacted states in Fig 6.53 shows none of the double peaked structure seen in InGaAs for the final state distribution of electrons. Fig. 6.41 showing the **k**-space distribution of secondary states indicates that the impacted states (corresponding to generated electrons) are singly peaked in **k**-space at Γ for low energy impacting holes. As was noted for the mean secondary carrier energies, plotted as the lightly shaded circles in Fig. 6.45, the generated electrons (Fig. 6.53) take a greater share of the available energy than the generated holes (Fig. 6.52).

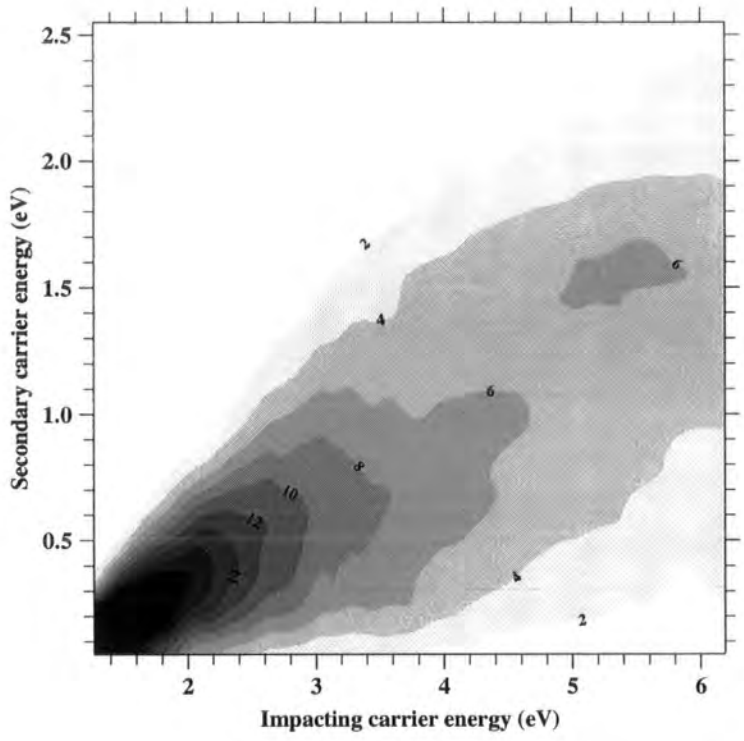


Figure 6.48: Distribution of final states (generated electrons) for initiating electrons in the 2nd conduction band of SiGe.

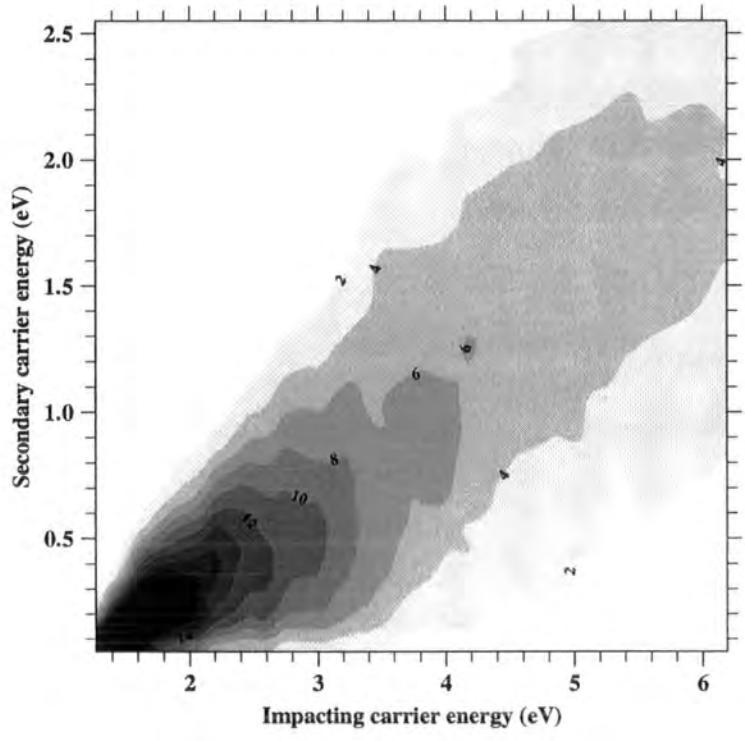


Figure 6.49: Distribution of impacted states (generated holes) for initiating electrons in the 2nd conduction band of SiGe.

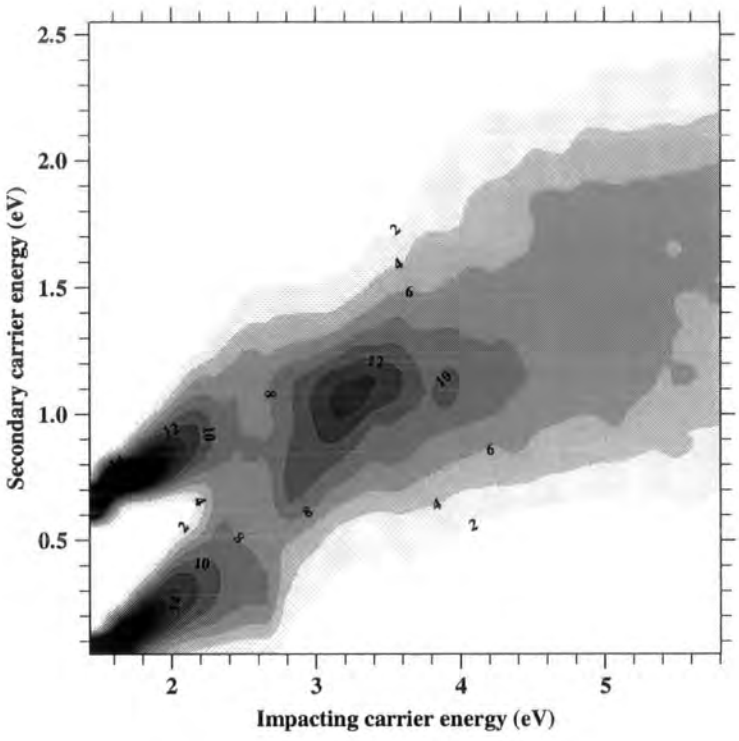


Figure 6.50: Distribution of final states (generated electrons) for initiating electrons in the 2nd conduction band of InGaAs.

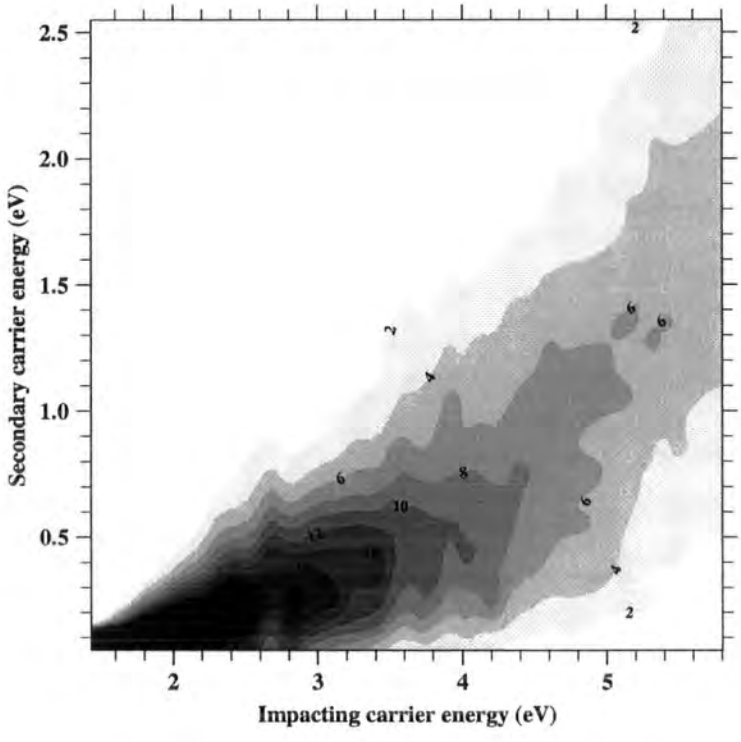


Figure 6.51: Distribution of impacted states (generated holes) for initiating electrons in the 2nd conduction band of InGaAs.

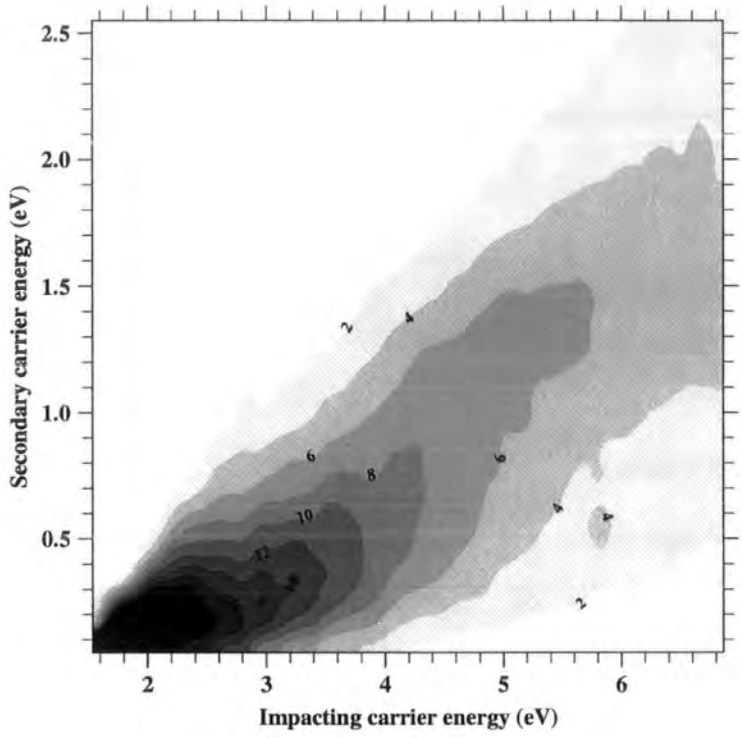


Figure 6.52: Distribution of final states (generated holes) for initiating holes in the spin split off band of GaAs.

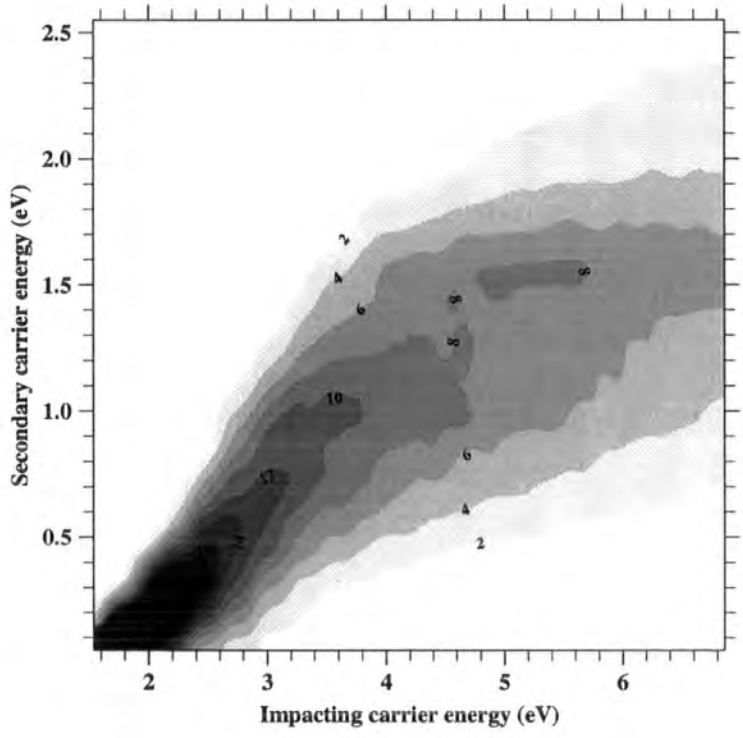


Figure 6.53: Distribution of impacted states (generated electrons) for initiating holes in the spin split off band of GaAs.

6.6 Comparison of Results with Other Authors

In this section, the results obtained in this work are compared to those obtained by other authors from calculations based on realistic band structure. Most authors integrate the rate using either Kane's method or a tetrahedron method, summarised below:

Kane's method Kane's method [58] is similar to the simple algorithm described in §5.1.1 of Chapter 5. Pairs of final states distributed uniformly in **k**-space throughout the Brillouin zone are considered. For each pair, the impacted state is chosen so as to conserve crystal momentum. Those which are then determined to also conserve energy to within a fixed tolerance (0.2 eV in the case of Kane) contribute to the total rate.

Tetrahedron method The Brillouin zone is discretised into tetrahedral sub-volumes. By interpolating energy and and matrix elements linearly within each sub-volume, a sub-volume's contribution to the rate can be determined analytically. The total rate is obtained by considering all sub-volumes in turn. Note that, within the approximation of linear interpolation, energy is conserved exactly.

The following is a list of the other authors whose work is compared, along with the method of integration used and some of the other features of their particular calculations:

Bude & Hess [20] The rate integration is performed for electrons in GaAs and InGaAs using a method similar to that of Kane, with local pseudopotential band structure. Matrix elements include the commonly neglected terms and a **q**- and ω-dependent ϵ.

Wang & Brennan [21] The rate for electrons in GaAs is integrated by a similar method to that of Kane, with the band structure provided by the $\mathbf{k} \cdot \mathbf{p}$ method.

The matrix elements are taken to be a statically screened Coulomb interaction in which ϵ is constant and the simple overlap approximation is used.

Kamakura, Mizuno, Yamaji, *et al* [22] The rate is calculated for electrons in Si using local pseudopotential band structure. Matrix elements are calculated including all the commonly neglected terms and $\mathbf{q}$- and ω-dependent ϵ. The rate integration is performed using a mesh of 2361 points throughout the Brillouin zone, spaced $\frac{1}{8}\left(\frac{2\pi}{a_0}\right)$ apart. Within each mesh cube the integral is performed by a tetrahedron method.

Jung, Taniguchi & Hamaguchi [26] The rate for electrons in GaAs is integrated by a tetrahedron method, with the band structure being provided by the local pseudopotential method. Matrix elements include a full $\mathbf{q}$- and ω-dependent dielectric function and the commonly neglected terms.

Oğuzman, Wang, Kolník & Brennan [27] The rate for holes in GaAs is integrated in the same way as by Wang and Brennan above, using $\mathbf{k} \cdot \mathbf{p}$ band structure. The matrix elements were calculated using the simple overlap approximation, assuming a statically screened Coulomb potential and $\mathbf{q}$-dependent ϵ.

Stobbe, Redmer & Schattke [59] The rate for electrons in GaAs is integrated using a method similar to that of Kane in which the delta function is approximated by a top-hat function 0.2 eV wide. The band structure is calculated using the local pseudopotential method, and the matrix elements are obtained from a statically screened Coulomb potential in which ϵ is taken to be $\mathbf{q}$-dependent and the commonly neglected terms are included.

Williams [60] The rate for electrons and holes in $Si_{0.5}Ge_{0.5}$ is integrated in a similar way to Kane, in which energy is conserved to within a given tolerance. The band structure is obtained using the local pseudopotential method, although it is fitted to correctly reproduce heterojunction band offsets, and as such the bulk

band gaps are inaccurate. Matrix elements are calculated using a **q**-dependent dielectric function and include the commonly neglected terms.

Stobbe, Könies, Redmer, Henk & Schattke [105] The rate for electrons in GaAs is integrated by using a special point method in which the delta function is approximated by a Lorentzian function whose FWHM is 0.4 eV. The band structure is obtained by the pseudopotential method, although the authors do not say whether local or non-local. Matrix elements are obtained from a statically screened Coulomb potential in which ϵ is constant and include the commonly neglected terms.

Sano & Yoshii [111] Band structure is obtained using a local pseudopotential calculation. The rate is calculated for electrons in GaAs and InGaAs by an approximate method which does not conserve **k** and treats the matrix elements as a constant which is fitted to give the rate calculated from first principles at an impacting carrier energy of 5 eV.

Fig. 6.54 shows the averaged electron initiated rate in GaAs compared with six other similar calculations. It can be seen that the the results of different authors at any given energy typically vary over about two orders of magnitude. The variations in rate are due to variations in the way the matrix elements are calculated, the differences between the band structures used (which in much of the other work under comparison here is obtained either by the $\mathbf{k} \cdot \mathbf{p}$ or local pseudopotential methods), and differences in the implementation of the numerical rate integration. From the figure it can be seen that the rates calculated in this work are among the lowest. Possible reasons for this are examined in §7.2 of Chapter 7.

For each line, an expression of the form $R = A(E - E_0)^P$ can be fitted, as described in §6.4.3. Table 6.8 lists the P parameters obtained for such fits for all the lines in Fig. 6.54.

Author	P
This work	5.2
JTH	6.7†
WB	2.3‡
SKRHS	0.8‡
BH	4.9
SY	5.6
SRS	4.0

Table 6.8: Comparison of the P parameters for electron initiated rates in GaAs. †The original paper fits the rate with two expressions: $P = 7.8$ ($E < 3.55$), $P = 5.6$ ($E > 3.55$). ‡The rate is poorly represented by the fit formula.

As discussed in §6.4.3, a higher value of P indicates greater deviation of the rate obtained using real band structure from that using idealised direct gap spherical parabolic bands and constant matrix elements (which give $P = 2$), and softer threshold behaviour. The fits marked with a ‡ symbol correspond to rates whose form is not well represented by the fitting formula, and thus the P parameter obtained may be misleading. The remaining fitted P-parameters are of roughly similar values, with the results of this work being fairly typical of them.

Fig. 6.55 compares hole initiated rates calculated here for GaAs with those of Oğuzman *et al* [27]. The agreement between the results is quite good, the rates calculated here being slightly lower as with the electron initiated rates. The P-parameters for each band can be examined in the same way as for the electron initiated rates, the results being tabulated below:

	This work	OWKB
SSO	3.2	4.4
LH	4.6	4.4
HH	4.4	2.0‡

Table 6.9: Comparison of the P parameters for hole initiated rates in GaAs. ‡The rate is poorly represented by the fit formula.

Clearly, results for the light hole bands are very similar, with a higher P-value (i.e. harder threshold) obtained in the spin split off band in this work. The rate in the heavy hole band obtained by Oğuzman *et al* is not well represented by the fitting formula, and so the value of P is misleading.

In Figs. 6.56 and 6.57, electronic rates in InGaAs and the rate for both types of carrier in SiGe are compared with the calculations of other authors. In both cases,

the rates calculated here are higher, and have a different type of dependence on the impacting carrier energy, the rates of the other authors appearing to be more exponential in nature. For the rates in InGaAs, the fitted P-parameters are 8.7 and 10.4 for Bude *et al* [20] and Sano *et al* [111] respectively. In SiGe the P-parameters obtained by Williams [60] are 13.4 and 9.5 for electrons and holes respectively. These very high values reflect the fact that the rates obtained by these authors have a dependence on energy that is closer to being exponential than of the form of Eq. (6.4).

In Fig. 6.58, the mean energies of final states obtained here and in the calculations of Jung *et al* [26] are compared. The energies correspond well between the plots. In addition, both plots show a similar spread of mean final state energies, and both show a slight kink in the dependence of the final state energies with respect to the impacting state energy at about 3 eV. Fig. 6.54 comparing the electronic rates in GaAs shows that the results of Jung *et al* are the closest to the results presented here over a wide range of energies.

Fig. 6.59 examines the accuracy with which energy is conserved in rate calculations performed here and by Kamakura *et al* [22]. Energy conservation requires that

$$E(\mathbf{k}_1) + E(\mathbf{k}_2) - E(\mathbf{k}_{1'}) - E(\mathbf{k}_{2'}) = 0 \tag{6.9}$$

where the energies are energy eigenvalues, rather than carrier energies. However, in the calculation performed here which explicitly relaxes the conservation imposed by the Dirac delta function, energy is conserved only to within half the width of the top-hat function used in the numerical integration. In addition, the Brillouin zone is discretised into cubes and energies interpolated linearly within these, which entails further inaccuracy in the energy conservation condition. The calculation of Kamakura also discretises the zone into cubic volumes. Within these, the rate is integrated by a tetrahedron method which, although it treats the energy conserving delta function exactly, nevertheless interpolates energies linearly. Thus in the calculation of Kamakura *et al*, any non-conservation of energy is due to interpolation errors alone.

The upper half of Fig. 6.59 plots the mean value of the left hand side of Eq. (6.9) against the energy of the impacting carrier for the calculations performed here and by Kamakura *et al.* Note that the calculation of Kamakura is for Si, whereas in this work the calculation is for SiGe. Nevertheless, energy conservation errors can be meaningfully compared. It is clear that energy conservation is more approximate in the calculation of Kamakura than in the calculation performed here, despite this work's explicit approximation of the energy conserving delta function. This is due to the different levels if discretisation used by the two calculations. Kamakura divides the zone into cubes of side length $\frac{1}{8}\left(\frac{2\pi}{a_0}\right)$, corresponding to 2361 **k**-points, which is fairly typical of the other calculations which require discretisation of the zone. In this work, the irreducible wedge is divided into cubes of side length $\frac{1}{64}\left(\frac{2\pi}{a_0}\right)$, which would correspond to more than 10^6 **k**-points throughout the zone. As a result, energies of the final and impacted states are interpolated more accurately here than by Kamakura. A more informative indication of the magnitude of the interpolation errors incurred by discretising the zone is given by the RMS energy conservation errors, plotted for this work in the lower half of Fig. 6.59. The error is typically around 10 meV, which includes the 2.5 meV error explicitly allowed by the top-hat function used here.

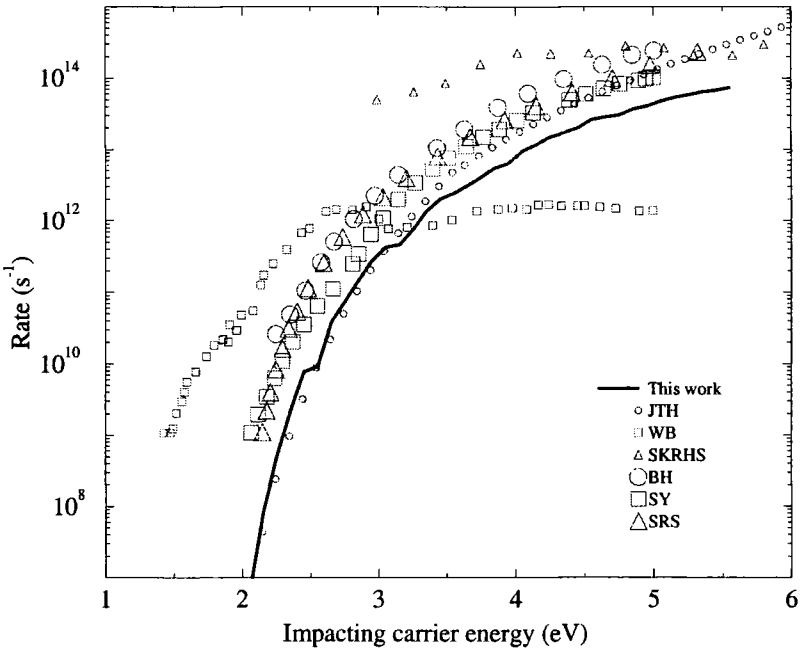


Figure 6.54: Electron initiated rates in GaAs. The other authors are: **JTH** Jung, Taniguchi & Hamaguchi; **WB** Wang & Brennan; **SKRHS** Stobbe, Könies, Redmer, Henk & Schattke; **BH** Bude & Hess; **SY** Sano & Yoshii; **SRS** Stobbe, Redmer & Schattke (see text for references).

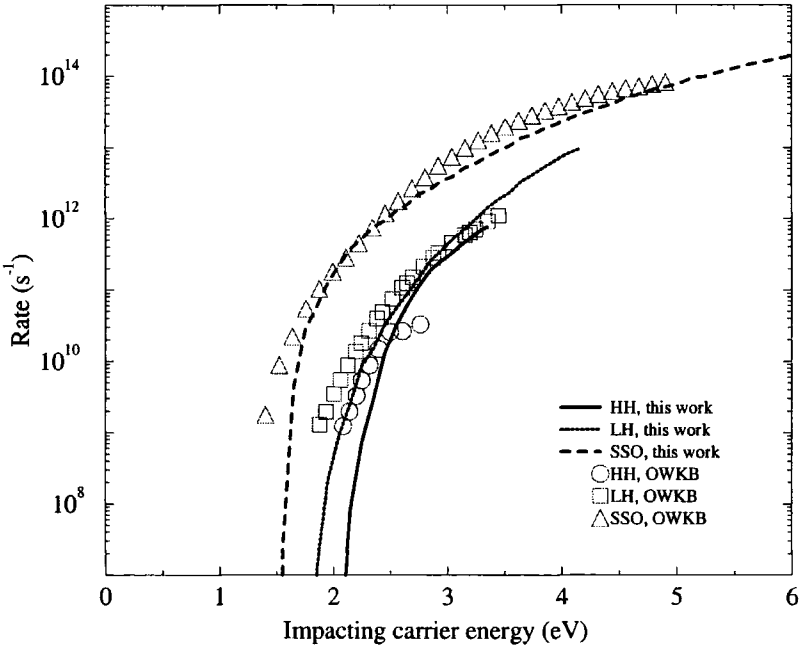


Figure 6.55: Hole initiated rates in GaAs. The other authors are Oğuzman, Wang, Kolník & Brennan.

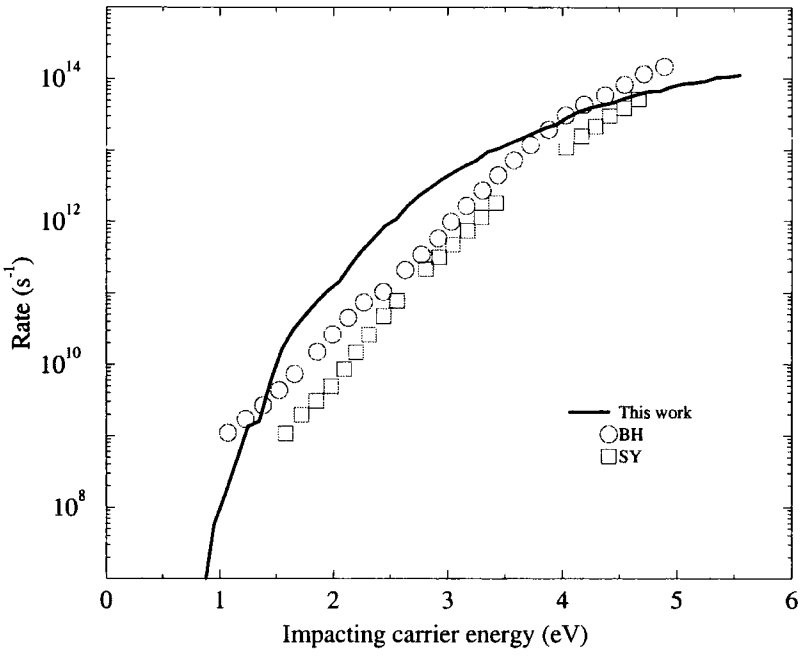


Figure 6.56: Electron initiated rates in InGaAs. The other authors are: **BH** Bude & Hess; **SY** Sano & Yoshii.

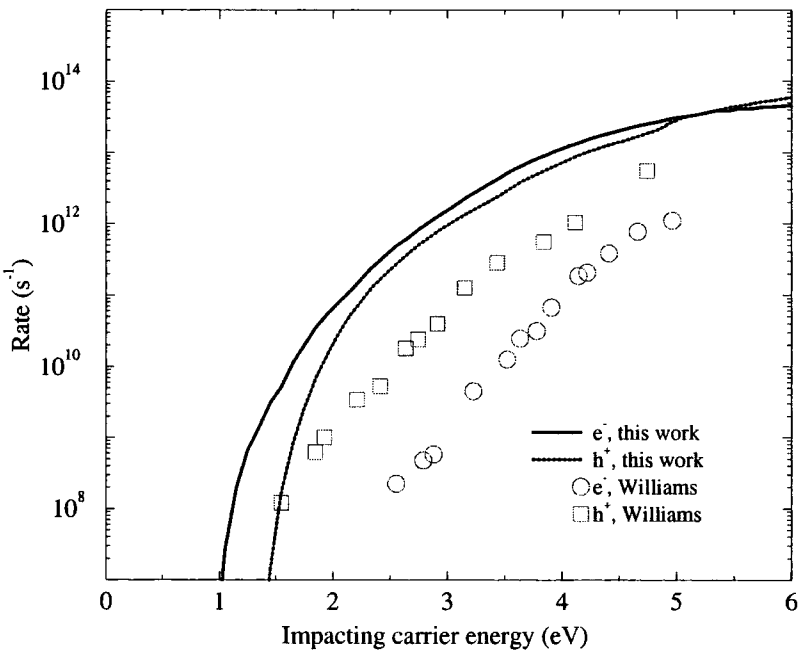


Figure 6.57: Electron and hole initiated rates in SiGe, compared to the results of Williams.

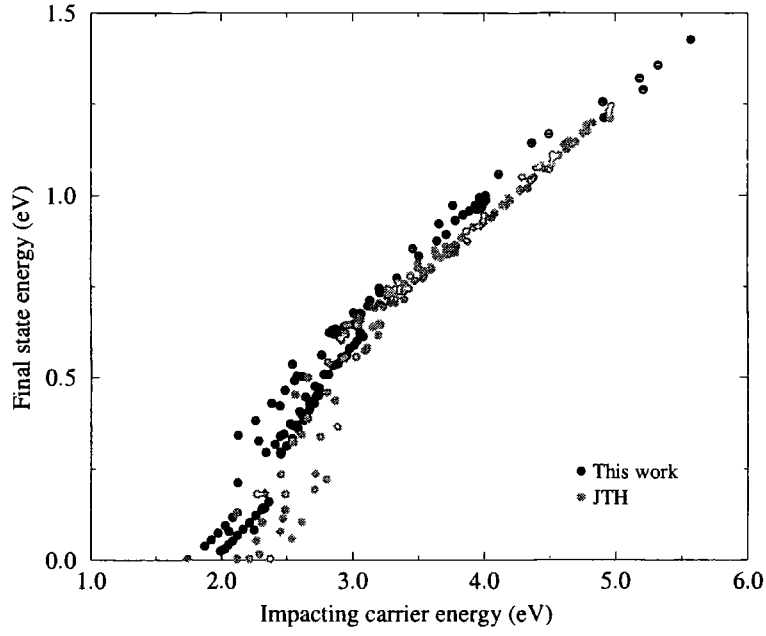


Figure 6.58: Mean energies of final state electrons generated by electron initiated transitions in GaAs. The other authors are Jung, Taniguchi & Hamaguchi.

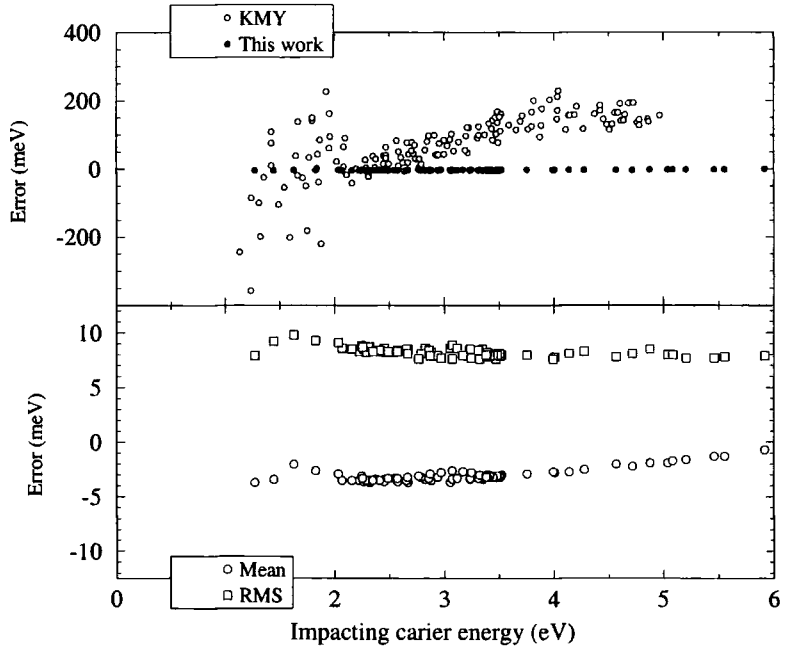


Figure 6.59: Errors in the conservation of energy incurred due to discretisation of the Brillouin zone. In the upper plot, the mean value of $E_1 + E_2 - E_{1'} - E_{2'}$ is plotted for this work and that of Kamakura, Mizuno, Yamaji, *et al.* In the lower plot, this work's results are re-plotted on a more suitable energy scale, along with RMS energy conservation errors.

Chapter 7

Analysis of Results

In this chapter, further analysis of the results is performed with the specific aim of understanding the main factors determining the magnitude of the rates in each material, and the causes of the qualitative differences in the properties of the materials.

7.1 Phase Space and Matrix Elements

The magnitude of the impact ionisation rate is determined by two factors: the area of the energy conserving surface in $\mathbf{k}_{1'}$,$\mathbf{k}_{2'}$-space, i.e. the volume of available phase space, and the average squared magnitude of matrix elements throughout this phase space. In this section the relative importance of each of these contributions is discussed.

In Fig. 7.1 the rate and the volume of available phase space are compared for initiating carriers in the second conduction band of InGaAs. In the upper half of the figure, the rate is plotted with respect to $\mathbf{k}$ as the dark lines (i.e. the same data as presented in §6.4.1). The quantities represented by the grey lines are calculated by setting the matrix elements to a constant, and describes the influence of phase space on the rate. The constant is chosen so as to make the volume of phase space give the best fit by least squares analysis to the actual rate. In the lower half of the plot, the mean value in arbitrary units of the squared magnitude of the matrix element is

plotted with respect to $\mathbf{k}$. This is obtained for a given impacting carrier simply by dividing the rate by the volume of available phase space.

Fig. 7.2 shows a similar plot for initiating carriers in the second conduction band of SiGe. Comparing Figs. 7.1 and 7.2, it can be seen that in each case the volume of phase space qualitatively reproduces the features of the rate. However, in SiGe there is good quantitative correspondence between the variation in the rate and the volume of phase space. This is due to the fact that the matrix elements in SiGe, plotted in the lower half of Fig. 7.2, do not in general vary as greatly with respect to $\mathbf{k}$ as in InGaAs. Note that where the rate is zero, the average matrix element is plotted as being zero, though is really undefined. The apparent rapid decrease in the matrix element occurring at $|\mathbf{k}| \simeq 0.7$ is not a real feature, but is due to an increase in statistical noise on the mean value of $|M|$ where the rate (and hence number of sampled points) is very low.

Similar behaviour can be seen in the hole initiated case. Figs. 7.3 and 7.4 compare rate and phase space plotted along the Γ–X line for the valence bands of InGaAs and SiGe. In InGaAs, the features of the rates are again approximately reproduced by the phase space. However, in each hole band the matrix elements have the effect of hardening the threshold (which is discussed further in §7.1.2), and in the light and heavy hole bands also disguise some of the structure in the dependence of the volume of phase space on $\mathbf{k}$ that can be seen around $|\mathbf{k}| = 0.5$. As was the case for electron initiated transitions, there is good quantitative agreement between the variation of phase space and rate in the valence band of SiGe. Here, as in the second conduction band, the variation of the average matrix element is much less in SiGe than in InGaAs. Matrix elements in GaAs, not plotted here, show similar behaviour to those of InGaAs, though with less variation in the valence band.

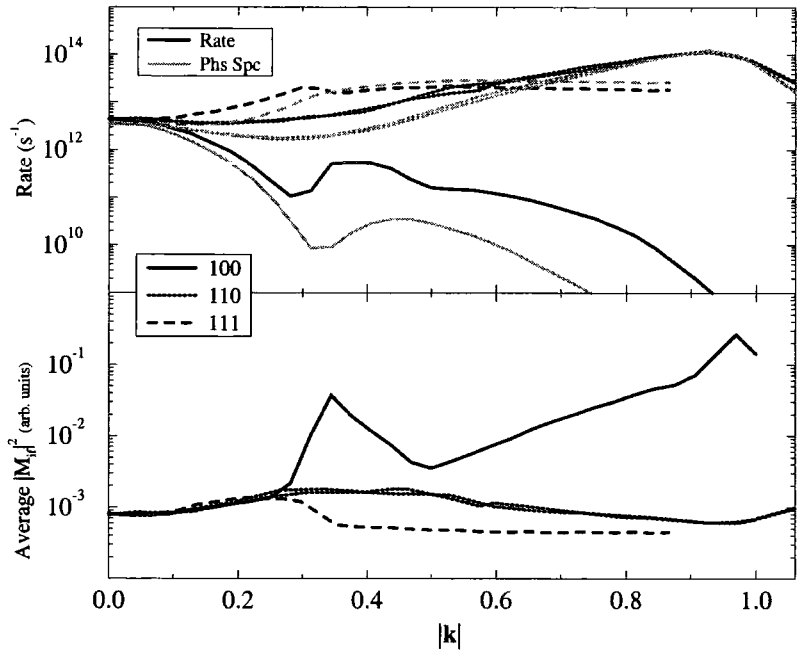


Figure 7.1: Comparison of rate and volume of available phase space, plotted with respect to **k**-vector of the initiating carrier along symmetry directions in the second conduction band of InGaAs

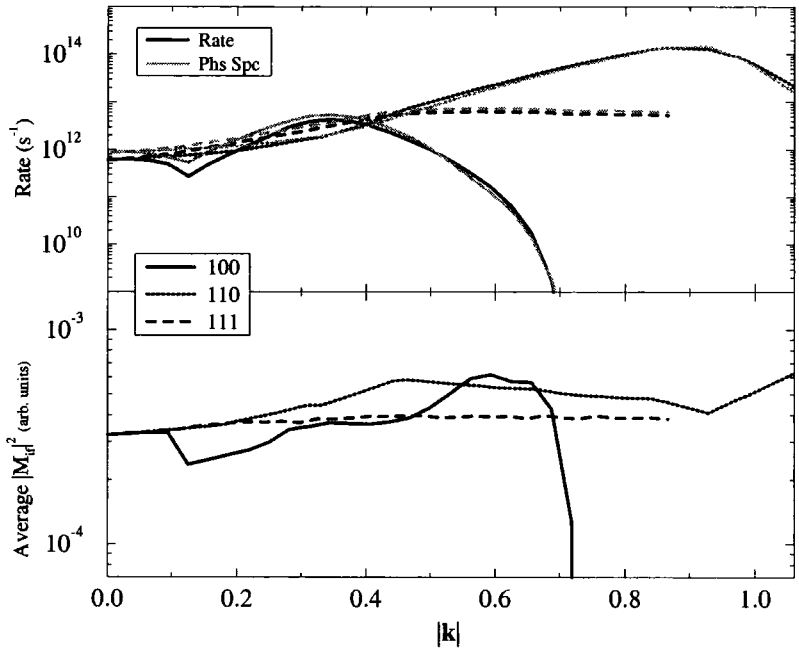


Figure 7.2: Comparison of rates and volume of available phase space, plotted with respect to **k**-vector of the initiating carrier along symmetry directions in the second conduction band of SiGe

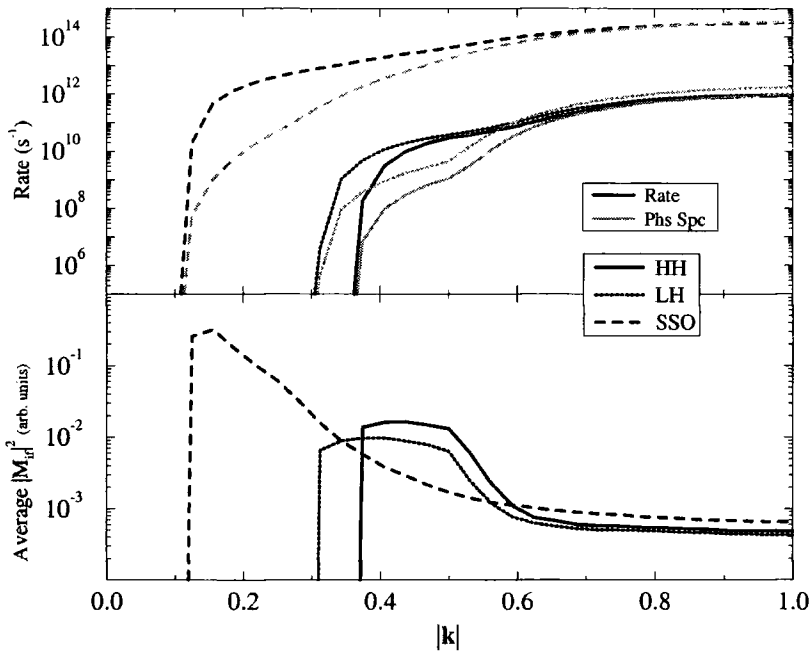


Figure 7.3: Comparison of rates and volume of available phase space, plotted with respect to **k**-vector of the initiating carrier along Γ–X in the valence bands of InGaAs.

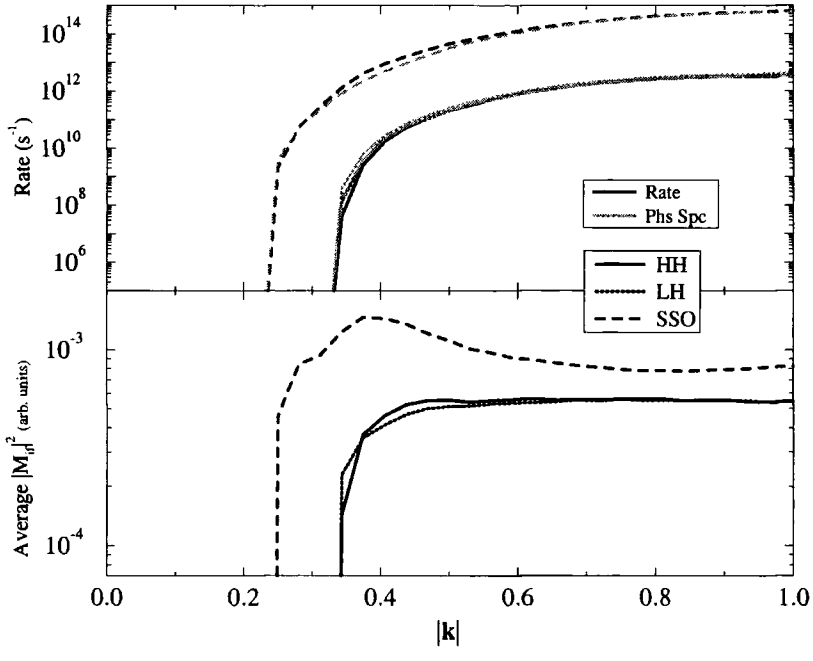


Figure 7.4: Comparison of rates and volume of available phase space, plotted with respect to **k**-vector of the initiating carrier along Γ–X in the valence bands of SiGe

7.1.1 Effect of Matrix Elements on Secondary State Distribution

The distribution of secondary states is determined by the shape of the energy conserving surfaces of allowed transitions. However, transitions to each possible final state occur with a probability proportional to the squared magnitude of the matrix element, and hence the distribution of secondary states will be further influenced by variations in the matrix element. Here, the importance of this latter effect is examined.

Figs. 7.5 and 7.6 indicate how the matrix elements affect the distribution of final states in InGaAs and SiGe respectively. The plots are of a similar type to those shown in §6.5.1. In each, the line graph at the top shows the rate due to transitions of the type CB2,HH→CB1,CB1 plotted with respect to impacting carrier wavevector along the line Γ–K. The row of surface plots immediately below shows the distribution of the final states in the first conduction band associated with impacting states located at the positions of the circles on the line graph. These final state distributions are calculated from the pairs of final states used in the Monte Carlo integration, each weighted by the corresponding matrix element. The lower row of surface plots also show the distributions obtained from the pairs of final states, but without including the weighting due to the matrix elements.

In the case of InGaAs, plotted in Fig. 7.5, each of the lower (unweighted) surface plots indicate that the final states are distributed throughout the Brillouin zone. At low impacting carrier energy, i.e. near Γ, they are generally confined towards the valley bottoms of the first conduction band while at higher energies they are located at all points of the zone. Comparing these distributions with those in the upper (weighted) row of plots, it is clear that the matrix elements have a significant effect on the distribution of final states in InGaAs. Specifically, the matrix elements act to favour the near vertical (low $\mathbf{q}$) transitions, as in each case the peak in the weighted distribution lies near to the position of the impacting state. Similar favouring of the low $\mathbf{q}$ transitions

is seen in GaAs (not shown here).

The final state distributions for SiGe, plotted in Fig. 7.6, show different behaviour. In SiGe, all X-valleys tend to be populated with final state carriers for impacting states of all wavevectors. The matrix elements tend to favour some valleys over others, but not strongly, and in particular do not act to favour the low **q** transitions. The weak influence of the matrix elements on the distribution of final states leads to the rate and phase space (plotted in Fig. 7.2) showing very similar variation as a function of the impacting carrier's wavevector.

The effect that the matrix elements have on the momentum transfer of transitions is examined more closely in Figs. 7.7 – 7.12. In each of these, the mean momentum transfer $\overline{q}$ of transitions for a given impacting carrier is plotted as a function of the carrier's wavevector. The solid lines show $\overline{q}$ calculated by weighting all transitions with the corresponding matrix element, while the dashed lines show the unweighted mean. (The horizontal dotted line at $\overline{q} \simeq 0.75$ shows the mean momentum transfer obtained by considering uniformly distributed random transitions).

The results for the first and second conduction bands and spin split off band of InGaAs are plotted in Figs. 7.7 – 7.9. In the first conduction band the mean **q**-transfer is approximately equal to the impacting carrier wavevector itself, as a result of the fact that transitions mainly occur to the Γ-valley. Furthermore, the similarity of the weighted and unweighted $\overline{q}$-values at all impacting vectors indicates that the matrix elements have little effect on the position of final states. This is to be expected, as in the first conduction band, where impacting energies are low, there is only a small range of possible final states and thus the matrix elements have little opportunity to affect the distribution. In the second conduction band, the matrix elements can be seen to have a much greater effect. For all impacting carrier wavevectors except those close to X (where the energy is lowest), the effect of the matrix elements is to reduce the momentum transfer, particularly near Γ. The effect of the matrix elements on transitions initiated by holes in the spin split off band is plotted in Fig. 7.9. As with

the first conduction band, the mean **q**-transfer is approximately equal to the impacting carrier wavevector due to final states lying near Γ, and similarly the matrix element generally has only a limited effect on this distribution.

Figs. 7.10 – 7.12 plot the mean momentum transfer as a function of impacting carrier wavevector for transitions in SiGe. Comparison of these plots with those for InGaAs shows three main differences. Firstly, the magnitude of the momentum transfer in SiGe is greater in all bands than in the corresponding bands of InGaAs. Secondly, the variation in $\bar{q}$ with respect to the impacting vector is much less in SiGe. Finally, the solid and dashed lines corresponding to any given crystallographic direction generally lie close together, indicating that the matrix elements do not affect the mean momentum transfer as they do in InGaAs, particularly in the second conduction band. A consequence of the different behaviour of the direct and indirect gap materials is examined in the following section.

Figure 7.5: The effect of the matrix element on final states in InGaAs. Impacting states lie along Γ–X in the 2nd conduction band. See also text on p.196.

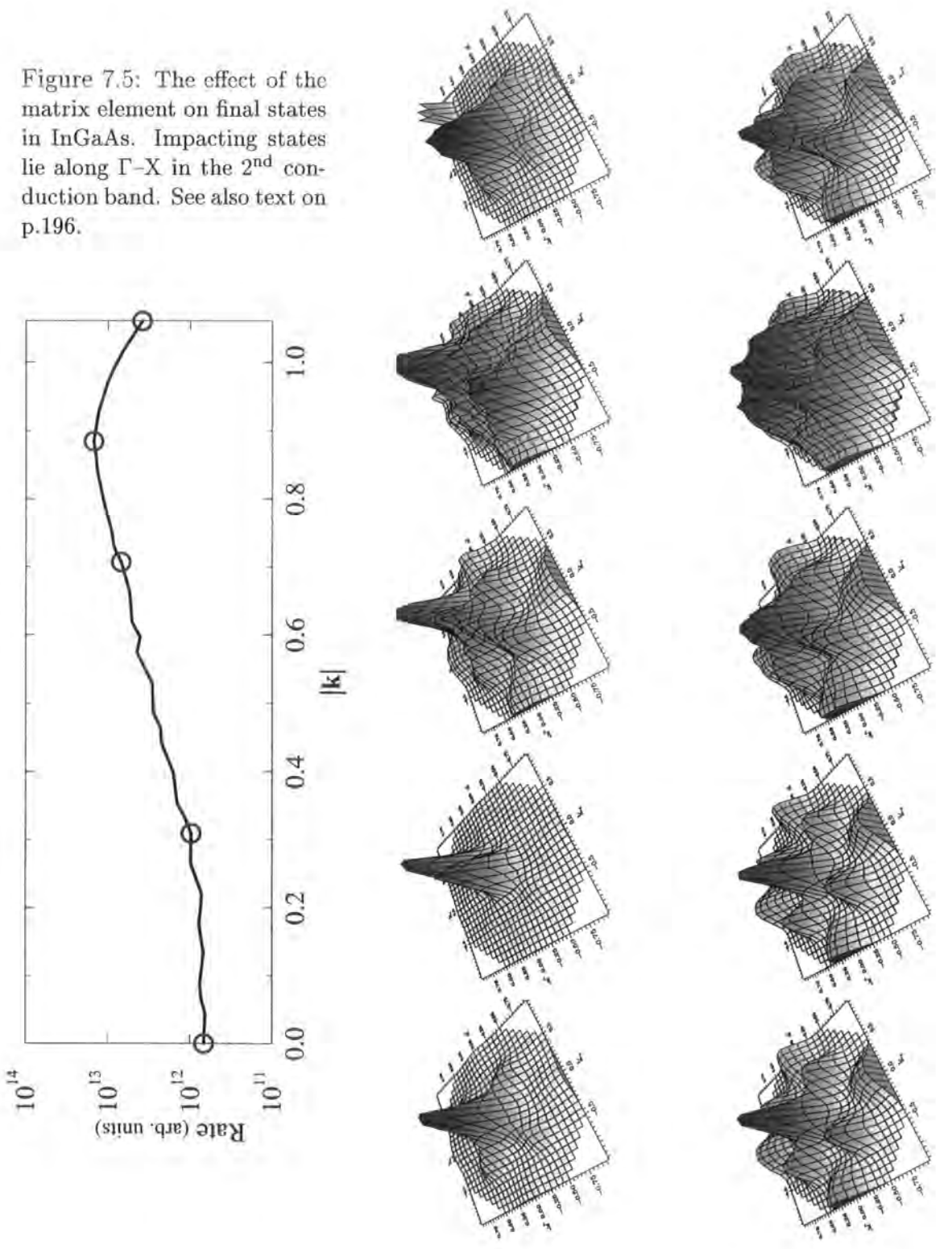
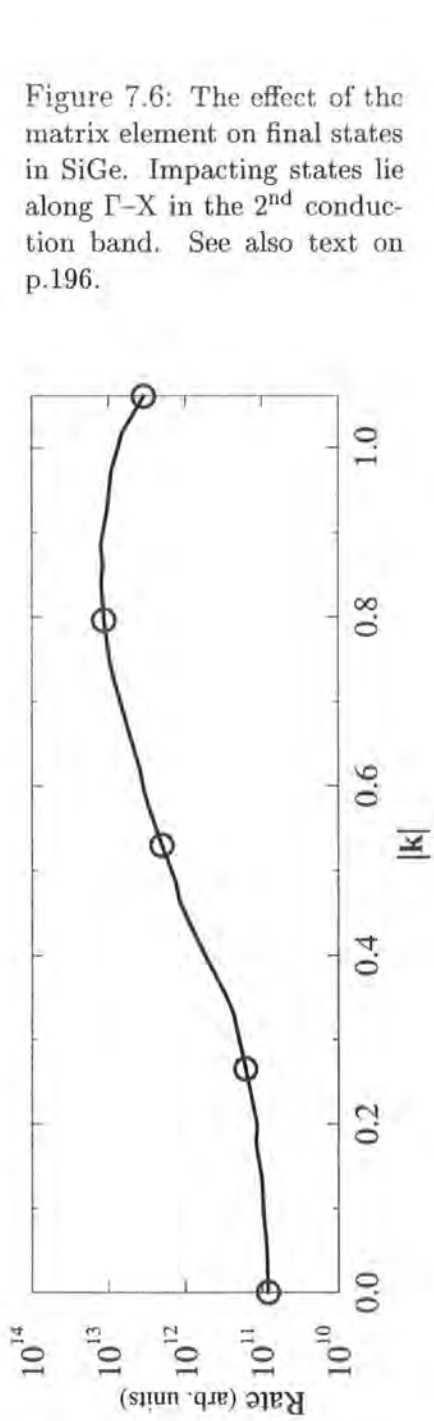
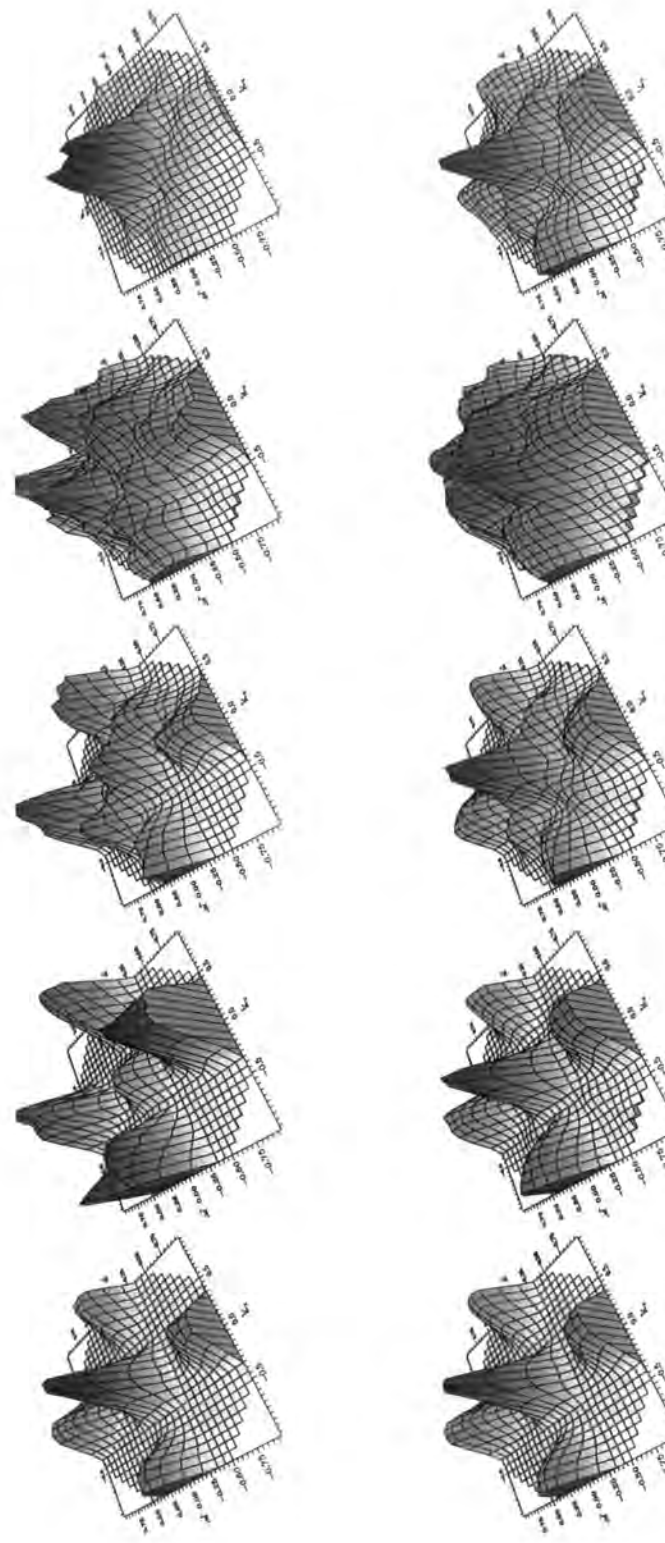


Figure 7.6: The effect of the matrix element on final states in SiGe. Impacting states lie along Γ–X in the 2nd conduction band. See also text on p.196.





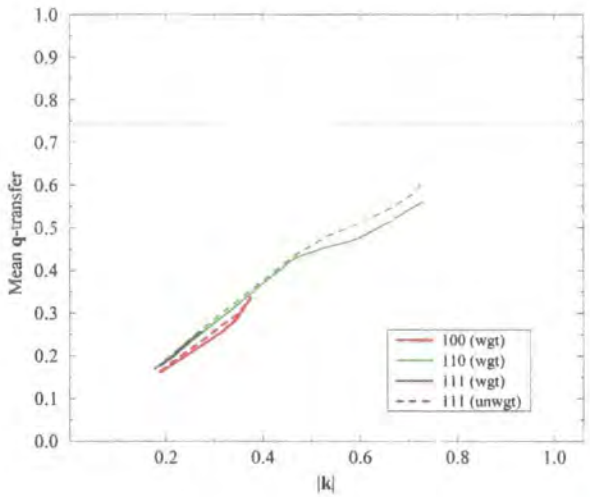


Figure 7.7: Mean **q**-transfer for transitions from the 1st conduction band in InGaAs, plotted with respect to **k**.

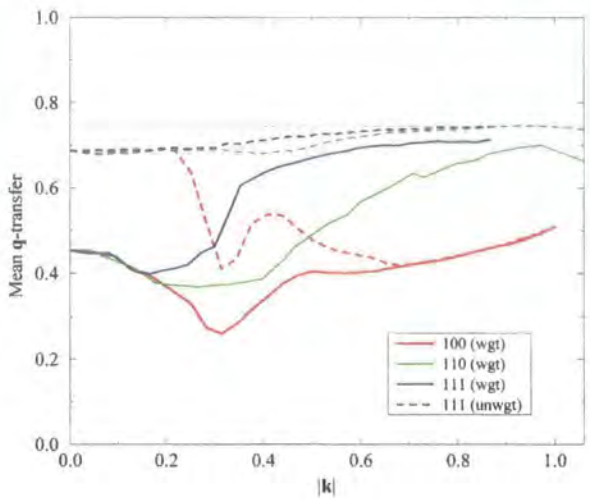


Figure 7.8: Mean **q**-transfer for transitions from the 2nd conduction band in InGaAs, plotted with respect to **k**.

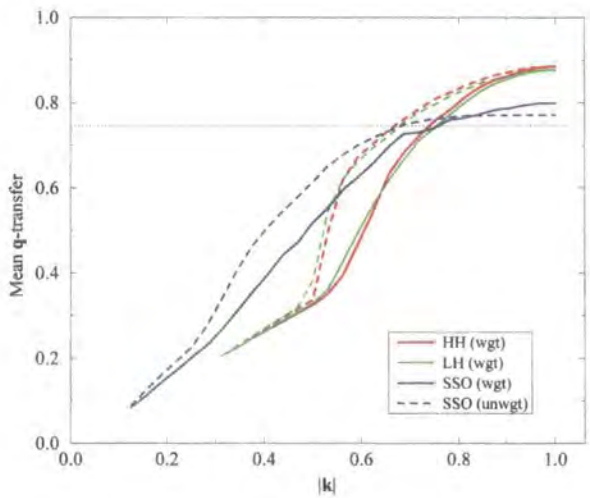


Figure 7.9: Mean **q**-transfer for transitions from the valence bands in InGaAs, plotted with respect to **k** along Γ–X.

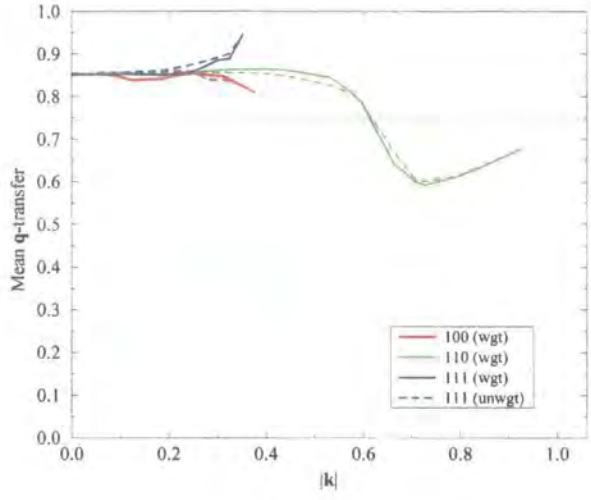


Figure 7.10: Mean **q**-transfer for transitions from the 1st conduction band in SiGe, plotted with respect to **k**.

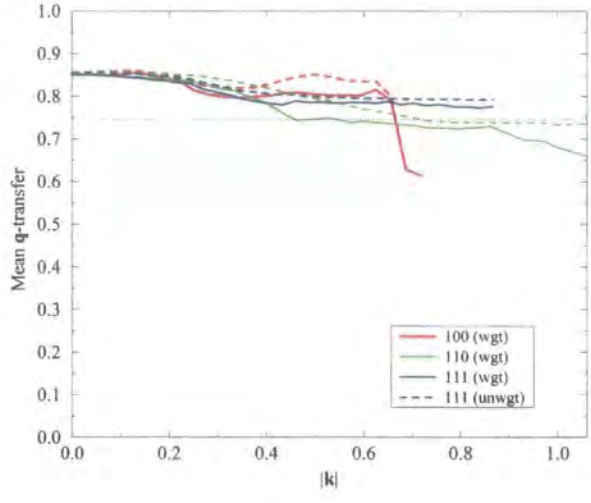


Figure 7.11: Mean **q**-transfer for transitions from the 2nd conduction band in SiGe, plotted with respect to **k**.

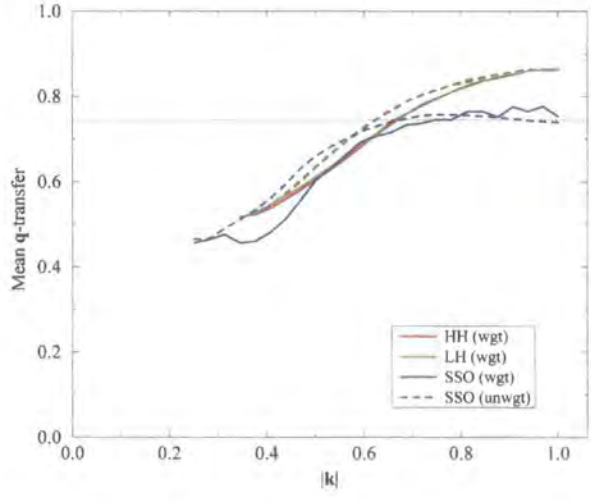


Figure 7.12: Mean **q**-transfer for transitions from the valence bands in SiGe, plotted with respect to **k** along Γ–X.

7.1.2 Effect of Matrix Elements on Threshold Softness

In §6.4.3 of Chapter 6, the calculated impact ionisation rates were fitted using the expression (repeated from Eq.(6.4))

$$R(E) = A(E - E_0)^P \tag{7.1}$$

where A, E_0 and P are the fitted parameters. As was noted in §6.4.3, the P parameter gives an indication of the softness of the threshold: a larger value of P indicates a softer threshold. The volume of available phase space can be fitted to the same expression as the rate, and the values of P thus obtained for different cases are compared in Table 7.1 to those for the rates themselves.

Band	Material					
	GaAs		InGaAs		SiGe	
	P_r	P_{ps}	P_r	P_{ps}	P_r	P_{ps}
SSO	3.2	5.5	2.6	4.9	3.5	3.9
LH	4.6	4.7	4.4	6.6	4.1	4.0
HH	4.4	4.3	5.4	6.3	5.2	4.7
CB 1	8.7	9.7	5.6	11.6	5.1	4.8
CB 2	4.7	5.5	4.3	6.9	4.1	4.1
e^-	5.2	6.1	5.6	9.4	4.9	4.8
h^+	5.1	6.1	4.2	6.4	4.7	4.4

Table 7.1: The P-parameter of Eq. (7.1) fitted to the rate, P_r, compared to that fitted to the corresponding volume of phase space, P_{ps}, for the various bands in each material.

Examining the values presented in the table shows that the rates tend to show harder threshold behaviour (that is, have a smaller value of P) than the corresponding volume of phase space in GaAs and InGaAs. In contrast, the rates tend to show slightly softer behaviour than the phase space in SiGe. This can be understood in terms of the momentum transfer data discussed in the previous section, in the following way.

As noted earlier in this chapter, the value of the impact ionisation rate can be considered to be the product of two factors: the available phase space and the average

matrix element. If the P parameter of the rate is lower than that of the phase space volume, as in the direct gap materials studied here, it follows that the average matrix element must be a decreasing function of impacting carrier energy. Similarly, in SiGe where the P-parameter of the rate is usually slightly higher than that of the phase space, the average matrix element must generally increase slightly with the impacting carrier energy. From Eq. (4.16) of Chapter 4 it can be seen that the squared magnitude of the momentum transfer appears in the denominator of the expression for the matrix element, and so we expect the dependence of $|M|^2$ on q to be given approximately by an expression of the form

$$|M_{if}|^2 \propto \frac{1}{q^4}. \tag{7.2}$$

Hence the relative behaviour of the rates and phase space volume in the materials studied could be explained if $\overline{q}$ were an increasing function of impacting carrier energy in GaAs and InGaAs and a decreasing function in SiGe. Fig. 7.13 plots the variation of $\overline{q}$ with impacting carrier energy for transitions initiated in the first conduction band of each material. The points represent the mean momentum transfer calculated for individual impacting **k**-vectors, while the solid lines are averages taken of these points. The wide spread of $\overline{q}$ values about the average indicates that the momentum transfer is not well represented as a function of energy alone, in common with the rates themselves. Nevertheless, trends in the value of $\overline{q}$ are clear from the graph: the mean momentum transfer is an increasing function of energy in the direct gap materials, and is relatively constant in the indirect gap material. This then leads to the matrix elements being decreasing functions of energy in the direct gap materials and relatively constant in the indirect gap material, as observed.

In Fig. 7.14, the variation of $\overline{q}$ with respect to impacting carrier energy is plotted for transitions initiated by holes in the spin split off band. In this case, $\overline{q}$ can clearly be seen to be an increasing function of impacting carrier energy in all three materials. This accounts for the fact that the P-parameter indicates a harder threshold for the

rates in the spin split off band than for the phase space in all materials, though much more so in GaAs and InGaAs, for which $\overline{q}$ increases more steeply with impacting carrier energy.

The data presented in this section suggests that a calculation which approximates the full expression for the matrix element with a fitted constant matrix element (i.e. calculations that approximate the rate by the volume of available phase space, suitably scaled) should expect to obtain softer electron thresholds for GaAs and InGaAs than would be obtained with a full calculation, and quite accurate (or very slightly harder) electron thresholds for SiGe. For hole initiated rates, the constant matrix element (CME) approximation predicts softer thresholds in the direct gap materials, in which the spin split off band dominates the total rate. In SiGe, the CME approximation also leads to a hole initiated threshold which is softer. However, no individual valence band dominates the rate in SiGe, and the relative positions of the threshold in each band is likely to be at least as influential on the characteristics of the overall rate as variation in the matrix element. Of the other authors whose work is compared in §6.6 of Chapter 6, Sano *et al* [111] used such a CME approximation, and did indeed obtain among the softest electron thresholds for GaAs and InGaAs (in their paper, they recognise the possibility of the CME approximation failing for direct gap materials). Figs. 7.15 and 7.16 compare electron initiated rates in InGaAs and SiGe obtained using the CME approximation with those obtained using the full matrix element. In each case the magnitude of the constant matrix element has been chosen so as to fit the phase space to the rate at the highest energy plotted. In the case of InGaAs, the CME approximation can be seen to badly underestimate the rate at low energy, while in SiGe it is a reasonably good approximation at all energies.

Although the analysis presented here has been applied to InGaAs and SiGe (results for GaAs, not presented here, show similar though less pronounced behaviour to those for InGaAs), the arguments based on **q**-transfer should be applicable in other materials. Thus we expect that it will generally be true that the use of the CME approximation

will lead to a softening of the thresholds for electron and hole initiated rates in direct gap materials, and little change in the threshold for electron initiated rates for indirect gap materials. It is difficult to make predictions about the threshold for hole initiated rates in the indirect gap case as it is likely to depend sensitively on the relative energies of the individual thresholds for the heavy, light and spin split off hole bands.

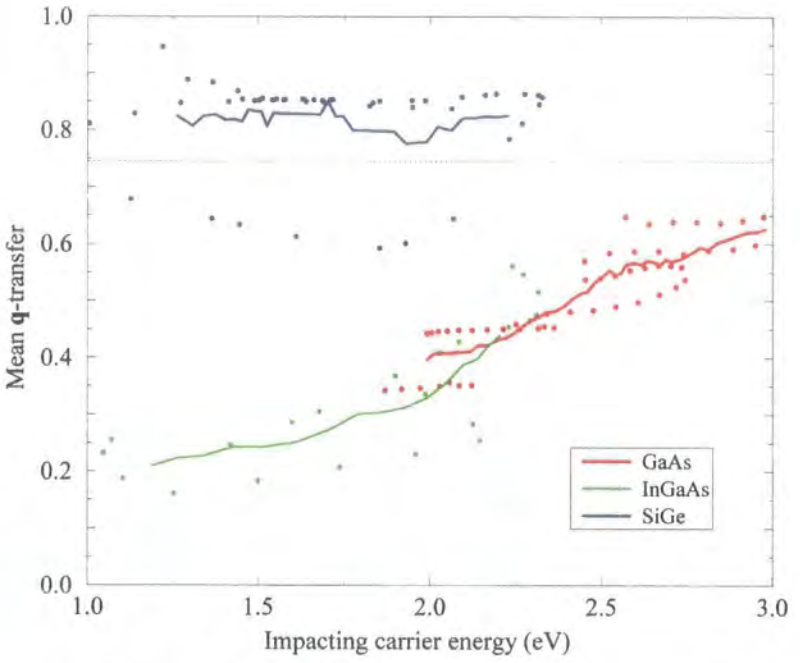


Figure 7.13: Mean **q**-transfer for transitions from the 1st conduction band, plotted with respect to impacting carrier energy.

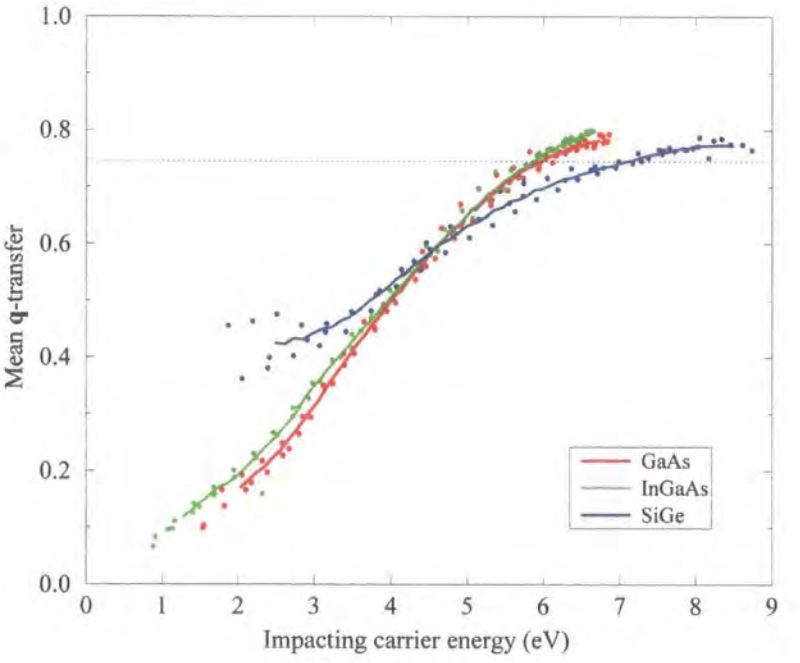


Figure 7.14: Mean **q**-transfer for transitions from the spin split off band, plotted with respect to impacting carrier energy.

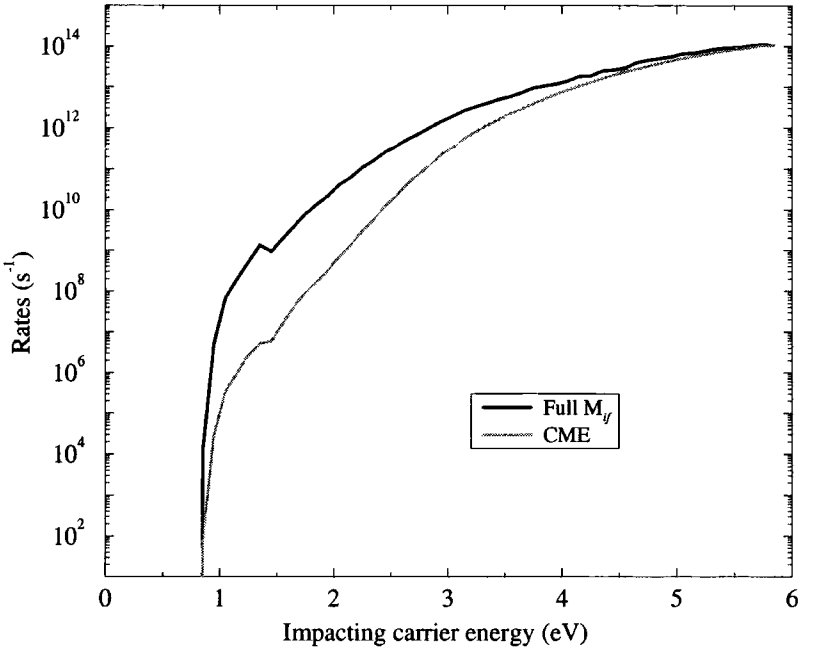


Figure 7.15: Electron initiated rates in InGaAs, calculated using the full matrix element, and the CME approximation.

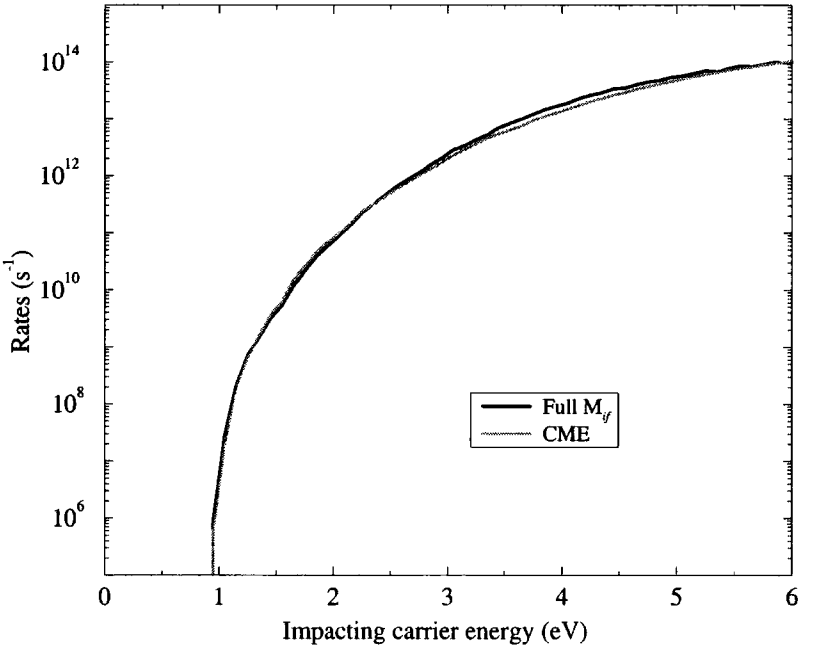


Figure 7.16: Electron initiated rates in SiGe, calculated using the full matrix element, and the CME approximation.

7.2 Approximations Made in the Rate Calculation

In §6.6 of Chapter 6, it was seen that for each material, there is considerable variation in the rates obtained by different authors. Causes of this variation arise from differences in:

- Approximations used in evaluating $|M_{if}|$. These include the number of plane waves used to expand the wavefunctions, the inclusion or neglect of the CNTs, and the form of the dielectric function used.
- Band structure. A number of methods are used to obtain electronic structure information (e.g. empirical pseudopotential, both local and non-local, and $\mathbf{k} \cdot \mathbf{p}$), and can involve fits to different sets of experimental data which may vary significantly for the same material. The band structure influences both the availability of phase space and the magnitude of the matrix elements.
- Numerical integration. Methods of integration vary, particularly with regard to the approximation of the energy conserving delta function and the degree of discretisation of the Brillouin zone.

This section examines the effect on the rates of variations in band structure, the inclusion or neglect of the CNTs, and the use of different approximations for the dielectric function. Note that the aim is not to determine the exact cause of the differences between each calculation, since this would be impossible without detailed knowledge each implementation, but to test whether the magnitude of the variation in rates produced by the use of different the approximations is sufficient to account for the variations seen between authors.

7.2.1 Effect of the Commonly Neglected Terms

The inclusion of the commonly neglected terms (CNTs) in the calculation of the matrix elements (as discussed in §4.2.1 of Chapter 4) requires more computational effort than

the use of the simple overlap approximation which is applicable to impact ionisation calculations for narrow band gap materials. Wang *et al* [21] and Stobbe, Könies, *et al* [105] are the only authors referred to in §6.6 that have performed electron initiated calculations neglecting the CNTs, and the rates they obtain show the least agreement with the general consensus of results obtained by the remaining authors plotted in Fig. 6.54. Oğuzman *et al* [27] have neglected the CNTs in calculating hole initiated rates in GaAs.

The CNTs have the greatest significance in matrix elements for which $\mathbf{q}$ is large, and so we expect to find that their inclusion is most important for electron initiated transitions in SiGe, and transitions initiated by holes near the Brillouin zone edge. Fig. 7.17 compares rates calculated for SiGe and InGaAs (which is assumed to be representative of GaAs also). As anticipated, they influence the rate most in the conduction band of SiGe, and at higher energies in the valence bands of both materials, increasing the rate by up to a factor of about five, and have a relatively small effect for most states in InGaAs. The rates for certain states above 4 eV in the second conduction bands of InGaAs and SiGe show anomalously large sensitivity to the inclusion of the CNTs. These carriers lie in a small volume of $\mathbf{k}$-space about the K-point of the Brillouin zone, and the relatively large difference in the results of the two calculations is due to a rapid fall in the value of the overlap integral $|\langle\psi_{\mathrm{CB2}}(\mathbf{k})|\psi_{\mathrm{CB1}}(\mathbf{k}{=}0)\rangle|^2$ as $|\mathbf{k}|$ increases from 0.9 to 1.0 along the Γ–K line, which is not seen in the full matrix element.

The magnitude of the difference in electron initiated rates in the direct gap material obtained by calculating the matrix elements with and without the CNTs is much too small to account for the large discrepancy in rates obtained by Wang *et al* [21] and Stobbe, Könies, *et al* [105] relative to the other workers whose results are plotted in Fig. 6.54. The discrepancy between the hole initiated rates in GaAs obtained here and by Oğuzman *et al* [27] is of the same order of magnitude as that introduced by the neglect of the CNTs. Calculations performed by Bude *et al* [20] and Williams [60], for InGaAs and SiGe respectively, included the CNTs, and so this factor cannot account

for the differences in rates obtained here and by these other authors. The calculation of Sano *et al* [111] for InGaAs used the CME approximation, making the inclusion or neglect of the CNTs irrelevant.

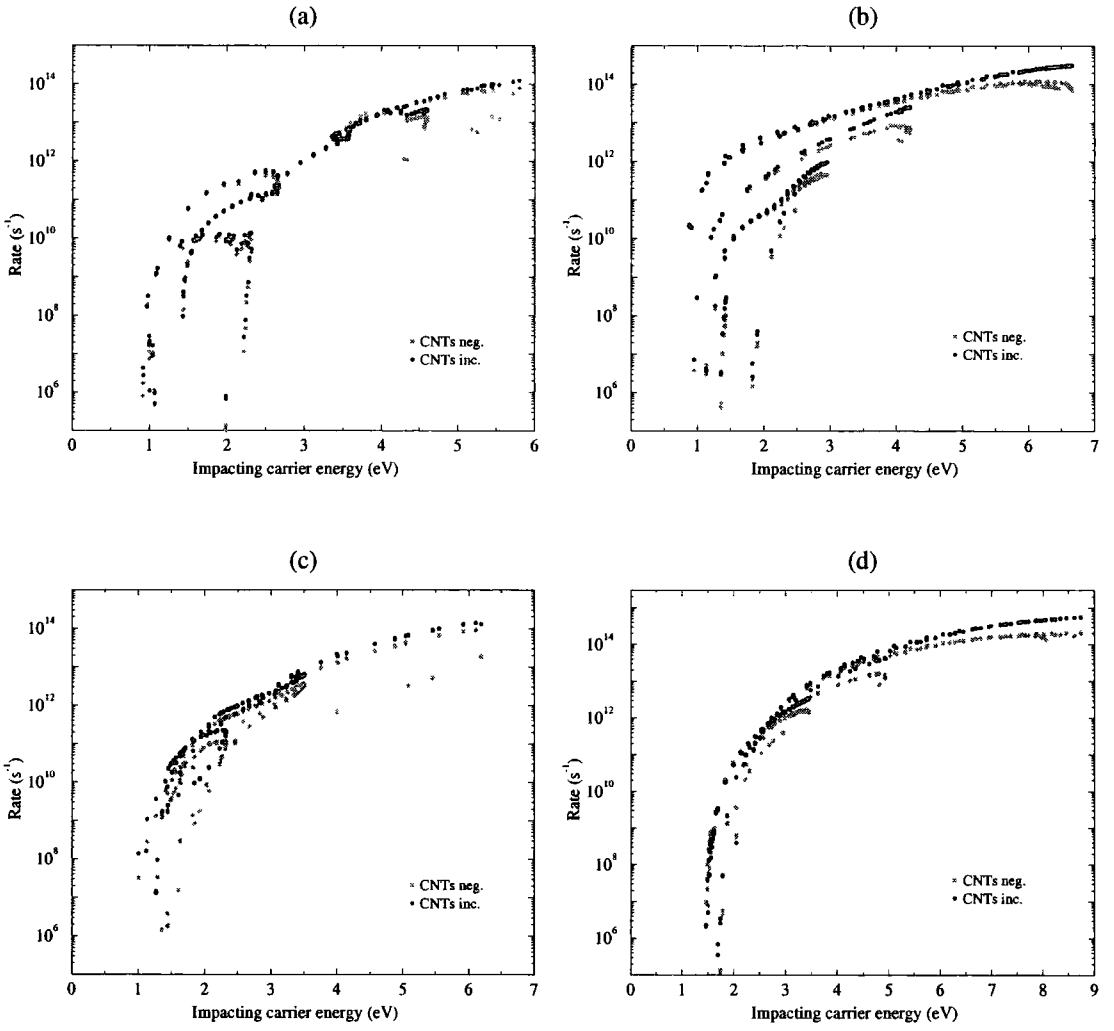


Figure 7.17: Comparison of rates calculated using the full expression for the matrix element (including the CNTs), and using the simple overlap approximation (neglecting the CNTs). Figs. **a** and **b** are rates in InGaAs for electrons and holes respectively. Figs. **c** and **d** are rates in SiGe.

7.2.2 Effect of the Dielectric Function

The functions used by the other authors cited in §6.6 to represent the dielectric response of the crystal can be divided into three types: q- and ω-dependent expressions (as used in this work), q-dependent expressions, and constants. To investigate the effects of these approximations, the q-dependent expression and constant were obtained using $\epsilon(q) = \epsilon(q, \omega{=}0)$ and $\epsilon_0 = \epsilon(q{=}0, \omega{=}0)$, where the general form $\epsilon(q, \omega)$ was that used elsewhere in this thesis.

The dielectric function appears in the denominator of the expression for the matrix element (Eq. (4.16) of Chapter 4), and so we expect $|M|^2$ and $\epsilon(q, \omega)$ to be approximately related by an expression of the form

$$|M|^2 \propto \frac{1}{|\epsilon(q, \omega)|^2}. \tag{7.3}$$

The variation in $|\epsilon(q, \omega)|^{-2}$ with respect to q and ω is plotted for InGaAs (which is typical of the other materials also) in Fig. 7.18. The plot indicates that the value of $|\epsilon|^{-2}$ is low at $(q = 0, \omega = 0)$, and generally highest at finite q-values along the line $\omega = 0$, with intermediate values at general (q, ω) (the rapid increase in $|\epsilon|^{-2}$ for $\hbar\omega \gtrsim 6$ lies beyond the range of energy transfer of interest here). Thus we would expect the rates obtained with each dielectric approximation to be lowest when using ϵ_0, intermediate for $\epsilon(q, \omega)$ and highest for $\epsilon(q)$. The rates plotted in Fig. 7.19 confirm these expectations.

The use of a constant expression for ϵ is found to be a poor approximation, particularly where **q**-transfer is high such as for impacting carriers lying near the zone edge in the valence bands of both materials, and for all impacting carriers in the conduction band of SiGe. The only authors referred to in §6.6 to use this approximation are Wang *et al* [21] and Stobbe, Könies, *et al* [105] (who have also neglected the CNTs). In the case relevant to their calculations, i.e. for electronic rates in a direct gap semiconductor, the use of the constant expression in place of a q- and ω-dependent one has the least effect and, as with the neglect of the CNTs, cannot account for the large discrepancy

between their rates and those of other authors.

Fig. 7.19 shows that the rates obtained using the $\epsilon(q)$ and $\epsilon(q,\omega)$ approximations for the dielectric function differ by a factor of up to about two for both materials and carrier types. The maximum discrepancy is seen where impacting carrier energy and hence energy transfer is greatest, as expected. This factor of up to two could account for differences in high energy hole initiated rates in GaAs obtained here and by Oğuzman *et al* [27], but is insufficient to account for the larger discrepancies between the electronic rates in GaAs and InGaAs, and rates for both types of carrier in SiGe obtained here and by authors using a q-dependent expression for ϵ.

Note that the q- and ω-dependent expression for the dielectric function used here is an isotropic approximation to the full $\mathbf{q}$- and ω-dependent expression given by Eq. (2.32) of Chapter 2. The RMS error introduced into the matrix elements due to the use of this isotropic approximation is estimated to be less than $\sim 5\%$, and the effect on the overall rate will be much less due to integration over many matrix elements.

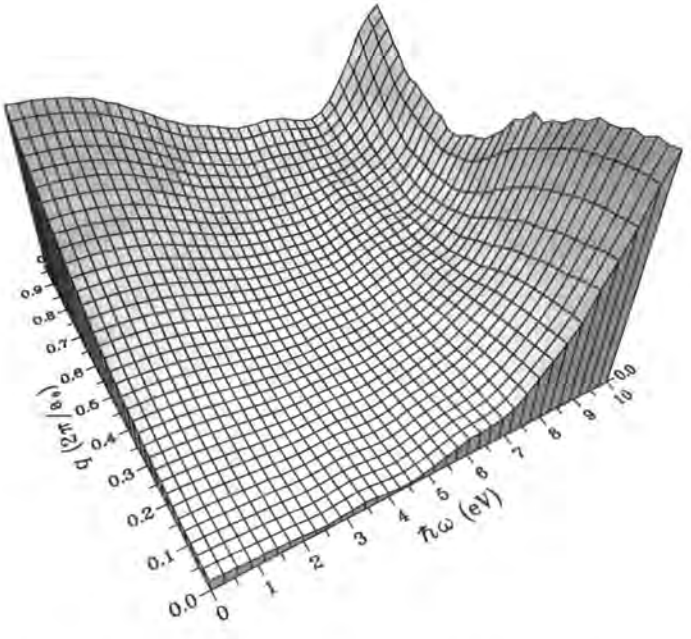


Figure 7.18: Variation of the function $|\epsilon(q,\omega)|^{-2}$ in InGaAs

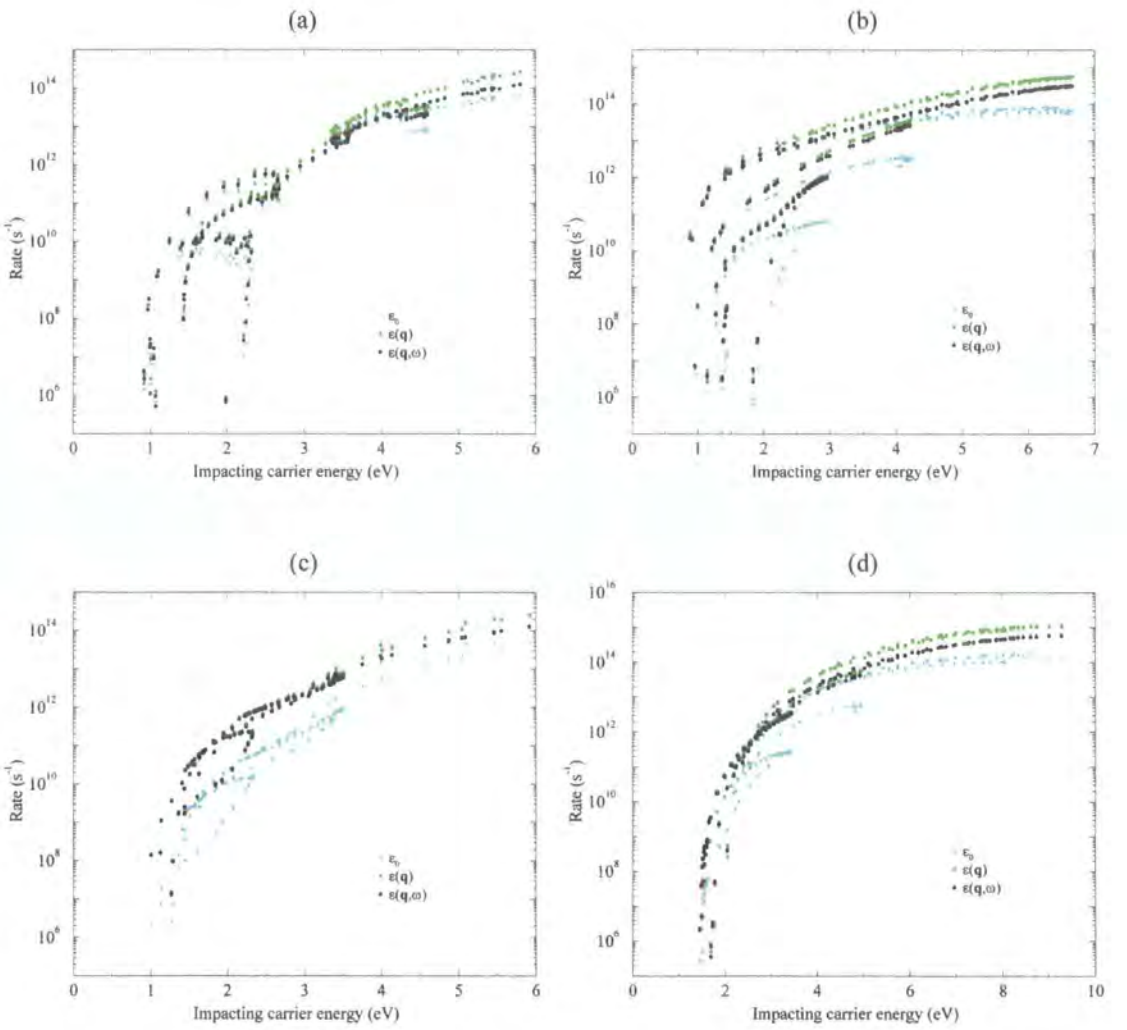


Figure 7.19: Comparison of rates using various dielectric function approximations. Figs. **a** and **b** are rates in InGaAs for electrons and holes respectively. Figs. **c** and **d** are rates in SiGe.

7.2.3 Effect of the Band Structure

The rate calculations compared in §6.6 use band structure obtained by various methods including local and non-local empirical pseudopotential methods and the $\mathbf{k} \cdot \mathbf{p}$ method, and the energy bands obtained by each will generally differ. In addition, the experimental data used to fit the band structure may differ from author to author, and even where the same data is used, the fitting procedure is itself not exact and different fits may lead to different calculated energy bands. Since the rates are sensitive to the exact shape of the bands, particularly at low energy, variation in the band structure is likely to be a factor contributing to the discrepancies between the calculated rates. In this section, rates in GaAs calculated using band structure obtained from two different pseudopotential calculations are compared. One is the calculation of Chelikowsky and Cohen [81], as used elsewhere in this thesis, which utilises a non-local pseudopotential and includes the effect of the spin-orbit interaction. The other calculation is that of Cohen and Bergstresser [89], which is based on a local pseudopotential method neglecting the spin-orbit interaction.

The energy bands calculated using the two methods are compared in Fig. 7.20. The overall form of the two band structures is broadly similar. However, in the local calculation, the energy of the X-valleys is lowered in comparison to the non-local case, the energies of the valence bands raised (i.e. carrier energies in the valence bands are lower in the local case) and the inclusion of the spin-orbit interaction in the non-local calculation splits the valence bands at the Γ-point. The dielectric functions at $q = 0$ obtained from the two calculations are compared in Fig. 7.21. The value of $\epsilon(q, \omega)$ is generally higher in the local case at energies of interest (although only the $q = 0$ case is plotted, this is also true for non-zero q-vectors).

Figs. 7.22 and 7.23 compare separately the contributions of the volume of available phase space and matrix elements calculated using the two band structures for electron and hole initiated transitions. For both types of carrier, the available phase space at any given impacting energy is larger when calculated using the local band structure. In

the case of electron initiated transitions, the lower X-valleys and higher valence bands both act to increase the number of possible final states. For the hole initiated transitions, it is perhaps surprising to find that despite the reduction in impacting carrier energy, the phase space volume is higher in the local case. However, the flatter valence bands obtained from the local calculation, while reducing the impacting carrier energy, correspondingly reduce the final state energy, making more final states accessible. In addition, the lower energy X-valleys (i.e. higher energy, in terms of holes) also provide more accessible impacted states.

From Eq. (7.3), we expect the larger value of $\epsilon(q,\omega)$ obtained from the local band structure to lead to matrix elements that are smaller by about 20%. Figs. 7.22 and 7.23 show that for both electrons and holes, matrix elements calculated using the local band structure are smaller typically by a factor of about two. The greater than expected reduction in matrix elements when going from the non-local to the local band structure is due to the different distributions of secondary states in each case. For example, in the case of electron initiated transitions, much of the increase in phase space volume is due to additional final states available in the X-valleys. However, matrix elements involving transitions to these states turn out to be smaller than for other transitions (in both band structures), and so the average value of $|M_{if}|^2$ is reduced.

Figs. 7.24 and 7.25 compare the rates[a] themselves obtained from the two band structures. The counteractive effects of increased volume of phase space and reduced matrix element in going from non-local to local pseudopotential calculation reduces the difference that might otherwise be seen in the rates obtained from these two band structures. Nevertheless, the increase in available phase space volume dominates and the rates for both types of carrier are higher when calculated using the local band structure. This would appear to account, at least to some extent, for the fact that rates calculated in this work are among the lowest of those plotted in Fig. 6.54.

[a]Note that the rates in Figs. 7.24 and 7.25 have been calculated for impacting carriers located along the 100, 110 and 111 directions only, and hence differ from the more complete calculations done for impacting carriers located throughout the zone, presented in Fig. 6.30.

The variation in rates obtained using the two band structures is greatest at low energy as expected and, in the case of GaAs, could account for much of the disagreement seen there in the results of other workers discussed in §6.6. The magnitude of the variation is not great enough to explain the much larger discrepancies between results previously reported and obtained here for InGaAs and SiGe. However, GaAs is a much studied material for which there is considerable experimental data, and hence the fitted band structures of different authors are likely to be similar (as in the case of Chelikowsky and Cohen's and Cohen and Bergstresser's fits). In other materials, particularly alloys, less data is generally available and the band structure may vary much more greatly between different calculations. Furthermore, in the case tested here, it turns out that the differences between the bands lead to changes in the surfaces of allowed transitions that just happen to lie at regions in which the matrix elements are significantly lower than their mean, diminishing the overall effect of the change. Stobbe *et al* [59] have investigated the effect of using different band structures in the calculation of the impact ionisation rate in GaAs, obtaining slightly greater variation in the rates than found here. Therefore, in InGaAs and SiGe, it is possible that variation in the band structure used by different authors has a greater effect on the rates.

A further source of variation in the rates obtained by different calculations is the number of plane waves used to expand the wavefunction when performing the pseudopotential calculation. This affects the shape of the energy bands, and the pseudowavefunctions and hence matrix elements. Convergence of the energy bands with respect to the number of plane waves used is rapid, being well converged when using 65 plane waves as in this work. The convergence of the matrix elements is discussed in §4.2.6 of Chapter 4. There it was shown that in InGaAs, the matrix elements are well converged when using 65 plane waves. In SiGe, convergence was found to be more of a problem, with errors in the matrix elements being of the order of 30%. However, these errors are negligible in comparison to the magnitude of the typical variation in rates obtained by the different authors discussed in §6.6.

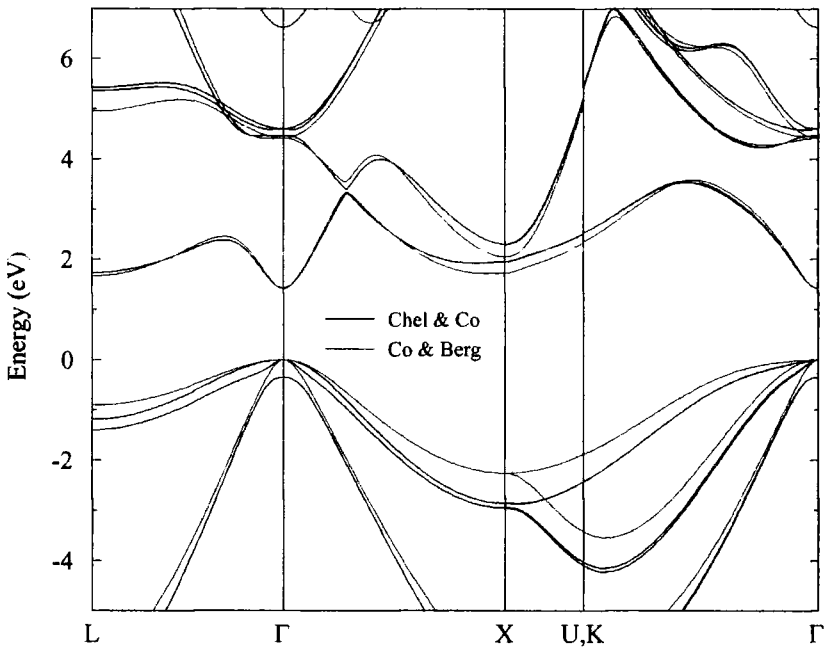


Figure 7.20: Comparison of the band structure of GaAs obtained using the non-local pseudopotential method of Chelikowsky and Cohen [81] and using the local pseudopotential method of Cohen and Bergstresser [89].

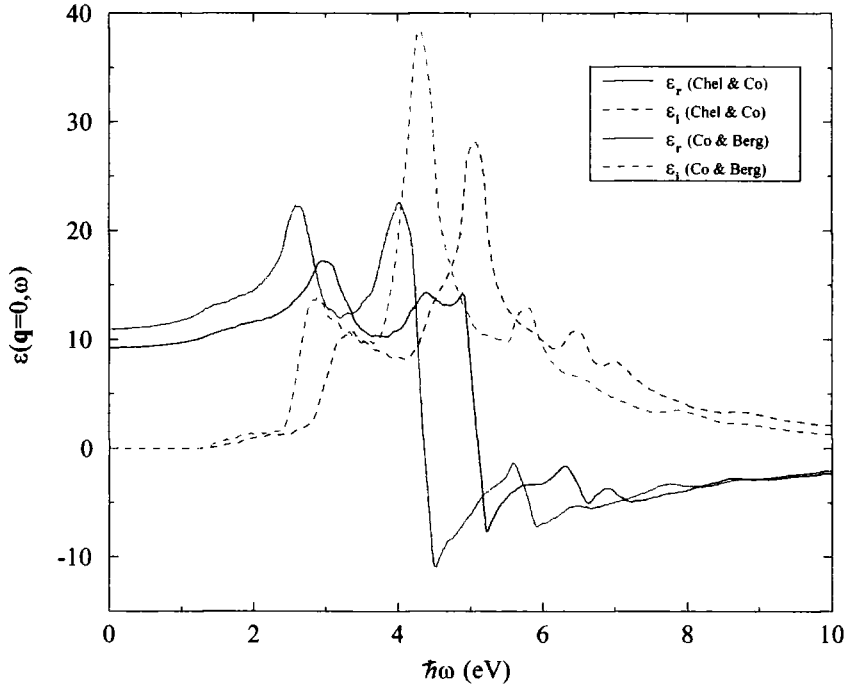


Figure 7.21: Comparison of the dielectric function of GaAs obtained using the non-local pseudopotential method of Chelikowsky and Cohen [81] and using the local pseudopotential method of Cohen and Bergstresser [89].

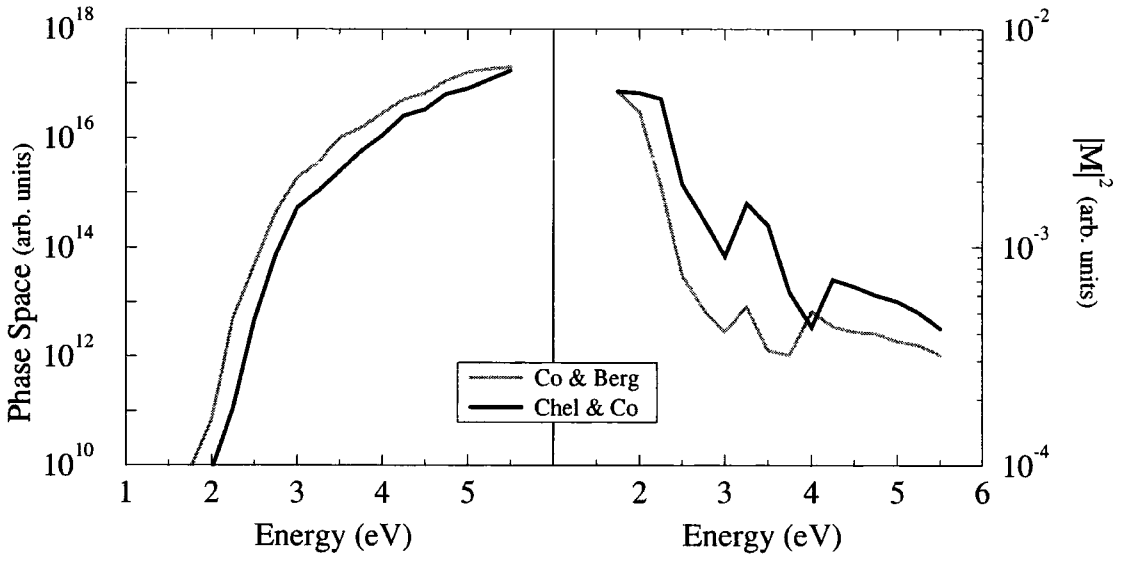


Figure 7.22: Average phase space and matrix element in GaAs plotted with respect to impacting electron energy, calculated using the local (Cohen and Bergstresser) and non-local (Chelikowsky and Cohen) band structures.

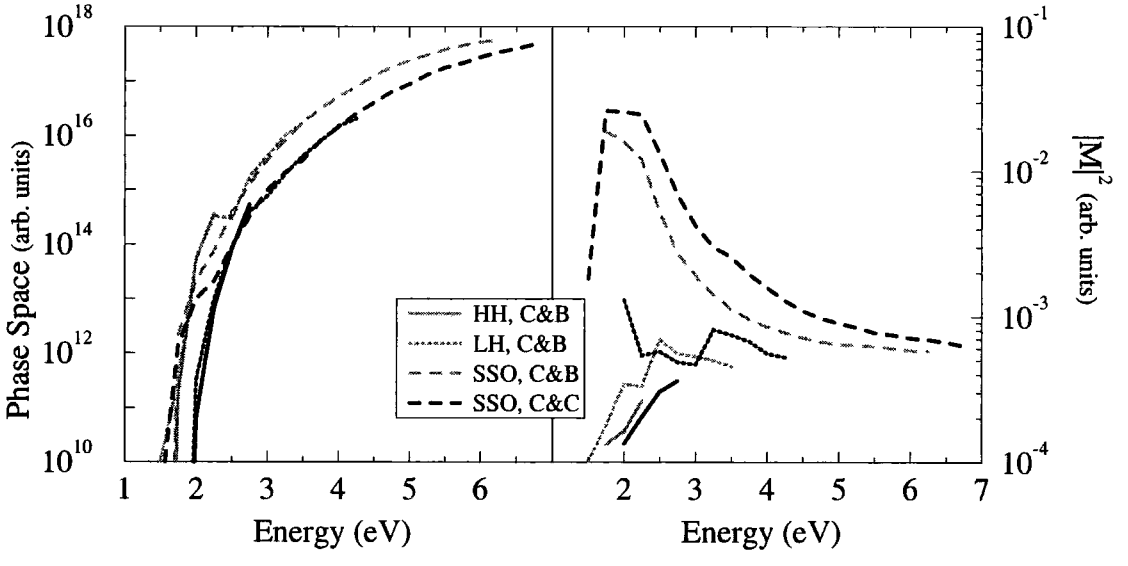


Figure 7.23: Average phase space and matrix elements in GaAs plotted with respect to impacting hole energy, calculated using the local (Cohen and Bergstresser) and non-local (Chelikowsky and Cohen) band structures.

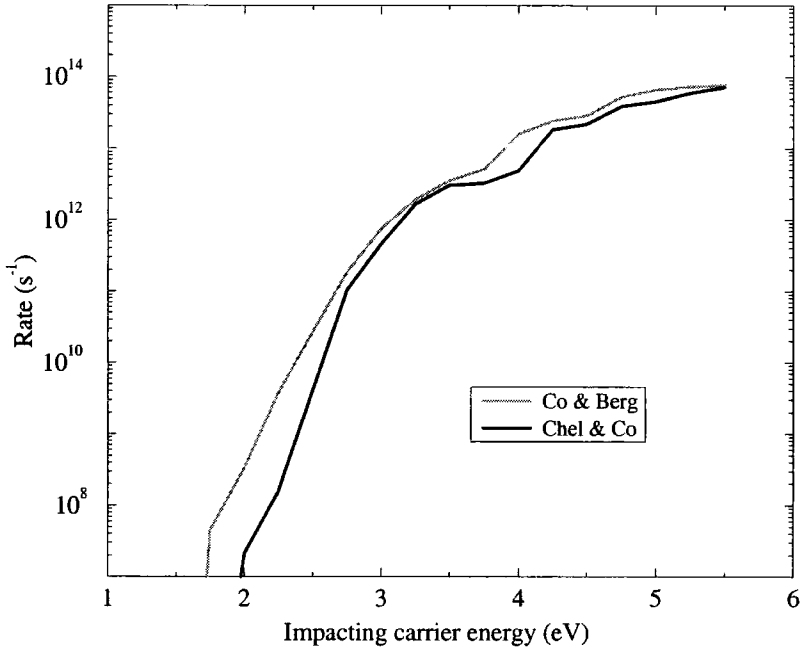


Figure 7.24: Electron initiated rates in GaAs, obtained using the local (Cohen and Bergstresser) and non-local (Chelikowsky and Cohen) band structures.

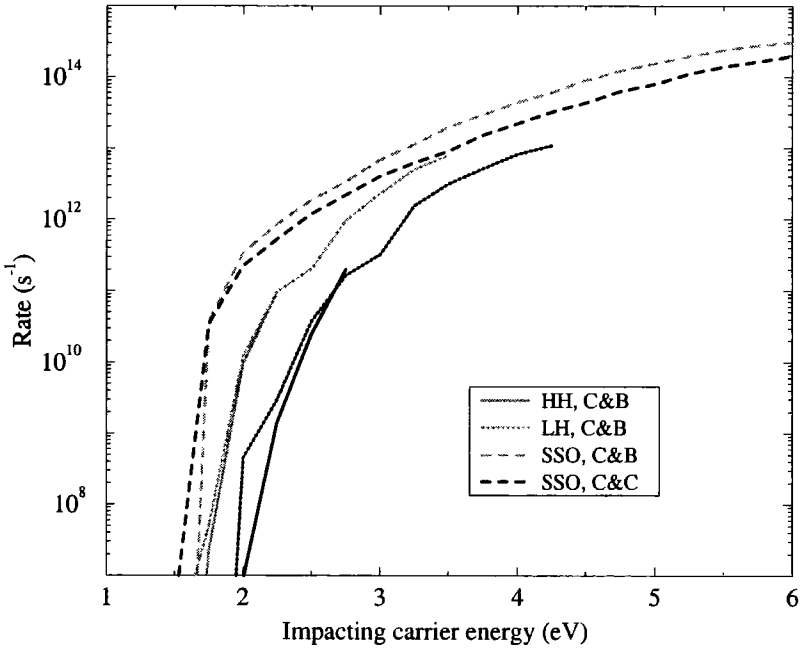


Figure 7.25: Hole initiated rates in GaAs, obtained using the local (Cohen and Bergstresser) and non-local (Chelikowsky and Cohen) band structures.

7.2.4 Variation in Rates: Summary

The neglect of the CNTs, and the use of a constant for the dielectric function are found to lead to significantly different rates from those obtained using the more sophisticated approximations. However, neither of these approximations is now widely used. Of the remaining sources of disagreement between authors, the use of a q-dependent dielectric function leads to over estimation of the rates at higher energy, in comparison to the rate obtained using a full q- and ω-dependent function, while uncertainty in the band structure can lead to variations in the predicted rate at lower energies. However, the wide variation in rates calculated by different authors cannot be fully accounted for by the factors discussed above, and so it must be assumed that differences in the implementation of the rate integration account for much of the variation in the rates.

7.3 The Importance of the Γ-Valley

A general feature of the results presented in this and the previous chapter is the fact that the direct gap materials studied frequently show qualitatively similar behaviour, while the indirect gap material has some distinctive properties. The origin of the differences is the existence in GaAs and InGaAs of a deep Γ-valley, which in SiGe is only very shallow and not the lowest part of the conduction band. The Γ-valley, having a light effective mass, does not provide a high density of states in comparison to the heavy effective mass satellite valleys, and it might therefore be expected that its influence on quantities involving integration over the Brillouin zone, such as the impact ionisation rate, would be small. In fact the qualitative differences in the properties of the direct and indirect gap materials studied here suggest that it is highly influential, and this is investigated in this section.

Fig. 7.26 shows the total density of states in the first conduction band of InGaAs, and the density of states lying in the Γ-valley[b]. For energies above 0.55 eV the satellite valleys quickly dominate the density of states. Fig. 6.46 of Chapter 6 shows that final states of this energy correspond to impacting electrons of carrier energy $\sim$ 2 eV. Thus, by considering only the density of states lying in the Γ-valley, we would expect this valley to have a small influence on the rates for impacting carriers above about 2 eV.

Fig. 7.27**a** shows electron initiated rates for InGaAs plotted as a function of the impacting carrier energy. The grey points represent total rates, calculated in the usual way, and the black points represent rates calculated by excluding all pairs of final states which include a state in the Γ-valley. For impacting carriers below about 2 eV, all transitions involve at least one state in the Γ-valley, since this valley provides the only final states of sufficiently low energy to be accessible. However, at higher energies, for which final states in the satellite valleys are available, the Γ-valley is still involved in more transitions than would be expected based on its contribution to the density

[b]A state is defined as lying in the Γ-valley if moving it down the energy gradient in **k**-space would take it to the Γ-point.

of states alone. Fig. 7.27**b** shows the total volume of available phase space (the grey points) and the volume of phase space calculated by excluding the Γ-valley (the black points). It is apparent that the influence of the Γ-valley on the total volume of available phase space is much less significant than on the rate.

The data in Fig. 7.27 is re-plotted in an alternative form in Fig. 7.28. The black points indicate the fraction of the total rate which is due to transitions involving at least one final state in the Γ-valley. The grey points similarly represent the Γ-valley's fractional contribution to the phase space. The dotted line is an estimate of the phase space provided by the Γ-valley based on the density of states of Fig. 7.26, and assuming the final state carrier energy E_f is related to the impacting state carrier energy E_i by

$$E_f = \frac{1}{3}(E_i - E_{gap}) \tag{7.4}$$

which is based on the simplifying assumption that the three generated carriers share the energy made available by the impacting carrier equally. From this plot it is clear that the fractional contribution of the Γ-valley to the total phase space drops off rapidly once the heavier satellite valleys become accessible, and is approximately equal to the value we would expect based on the 3-dimensional density of states in this valley. However the fractional contribution to the rate remains much higher, accounting for the majority of transitions for impacting carriers up to about 3.5 eV. The fact that the Γ-valley has a greater significance in influencing the total rate than would be expected from its contribution to the phase space indicates that the corresponding matrix elements are higher. This is due to the low **q**-transfer involved in transitions from the top of the valence band to states in the conduction band near Γ.

By excluding transitions involving the Γ-valley, the properties of the rates calculated for the direct gap materials GaAs and InGaAs become similar to those calculated for the indirect gap SiGe. Fig. 7.29 compares electron initiated rates in the three materials. In Fig. 7.29**a** the total rates are plotted, calculated in the usual way. In Fig. 7.29**b**, the rates have been calculated by excluding final states in the Γ-valley. The

highly **k**-dependent nature of the total rates at low energy in the direct gap materials, particularly InGaAs, is reduced when the Γ-valley is excluded, and the rates become relatively well represented by a function of energy alone, as it was noted in §6.4.2 of Chapter 6 is already the case in SiGe.

Fig. 7.30 compares the electron initiated rate in InGaAs calculated using the full expression for the matrix element, and using the CME approximation discussed in §7.1.2. In Fig. 7.30**a** the total rates are plotted[c], and in Fig. 7.30**b** the rates obtained by excluding final states in the Γ-valley are plotted. In §7.1.2 it was noted that the CME approximation leads to softer thresholds when applied to the direct gap materials, but is a good approximation in the indirect gap. From Fig. 7.30 it can be seen that when the Γ-valley is excluded, the accuracy of the CME approximation is much improved for the direct gap material, and in this case InGaAs shows similar behaviour to SiGe. Table 7.2 compares values of the P parameters obtained by fitting Eq. (7.1) to the rates calculated with the Γ-valley included or excluded. It shows that the A and P parameters for each material are quite similar for the fits in which the Γ-valley has been excluded, i.e. all three materials show similar threshold hardness in this case. Fits obtained for the volume of phase space show correspondingly good agreement between materials when the Γ-valley is excluded, indicating that the closeness of the behaviour seen for each of the materials is due to genuine similarities and not merely to different dependencies of the volume of phase space and magnitude of the matrix elements leading to coincidentally equivalent results for the rates.

In GaAs, the Γ-valley plays a similar role in softening the thresholds, though the effect is less marked than in InGaAs due to the smaller separation of the Γ and satellite valleys in GaAs. Hence the P-parameter for electron initiated rates in GaAs lies between that of SiGe and InGaAs. Bude and Hess [20] have also noted that thresholds are expected to be softer in materials for which the Γ-satellite valley separation is larger

[c]Note that although Figs. 7.15 and 7.30**a** are equivalent plots, they differ quantitatively due to the fact that the rates in Fig. 7.30 were calculated using a reduced (but representative) set of impacted and final state bands.

and the band gap smaller, and Allam [67] has noted that we should expect enhanced probability of transitions to the Γ-valley due to the low **q**-transfer involved. Allam also argued that although the materials Si, GaAs, InAs and $In_{0.53}Ga_{0.47}As$ have a wide range of band gaps, they are all similar in that they have similar values of $\langle E_{ind} \rangle$, defined as

$$\langle E_{ind} \rangle = \frac{1}{8}\Big(E_\Gamma + 3E_X + 4E_L \Big) \tag{7.5}$$

where E_V is the energy gap between the top of the valence band and the conduction band valley located at V, and therefore similar rates are obtained in each material at high impacting carrier energy. Thus the similarities obtained here between materials when the Γ-valley is excluded may not extend to InP for example, which has a rather larger value of $\langle E_{ind} \rangle$.

	Γ included			Γ excluded		
	A	P	E_0	A	P	E_0
GaAs	3.1×10^{9}	6.5	1.64	2.5×10^{10}	4.8	2.14
InGaAs	1.1×10^{8}	7.0	0.17	3.6×10^{10}	4.5	2.04
SiGe	1.4×10^{10}	4.4	0.87	1.4×10^{10}	4.4	0.87

Table 7.2: Fitting parameters for electron initiated rates calculated by considering all possible final states, and by excluding final states in the Γ-valley. Note that parameters on the left hand side of this table are fitted to impacting vectors located along the 100, 110 and 111-directions and hence do not match those listed in Table 6.7 for carriers throughout the zone.

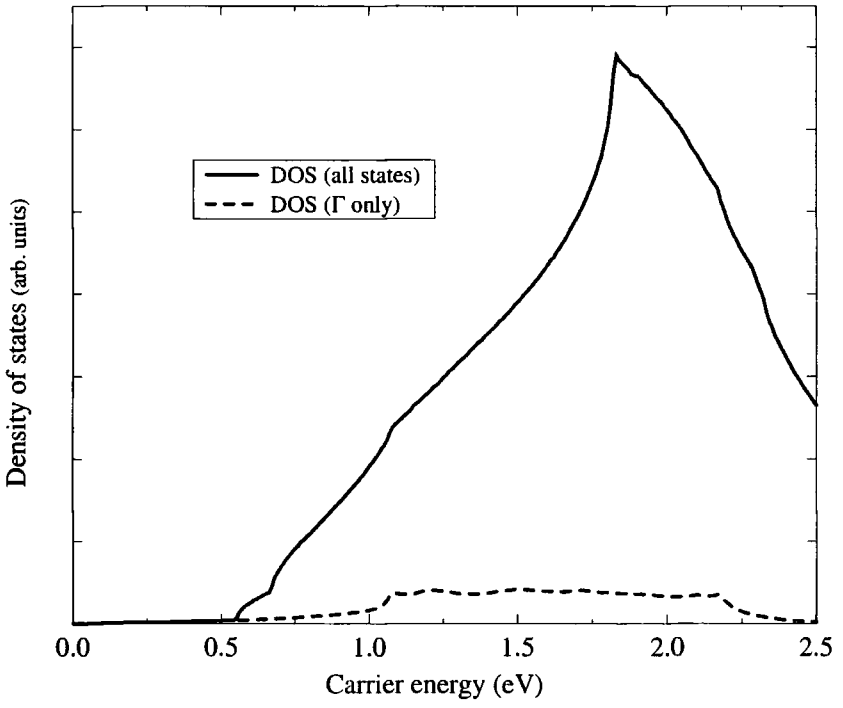


Figure 7.26: Density of states in the 1st conduction band of InGaAs. The solid line is the total D.O.S. while the dashed line is the D.O.S. in the Γ-valley only.

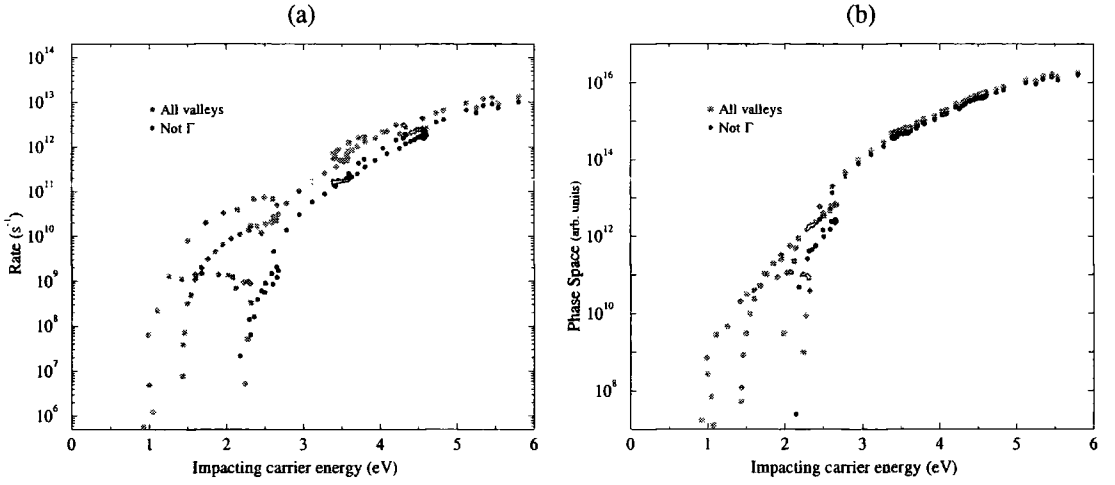


Figure 7.27: Comparison of rates and phase space for InGaAs evaluated by including all transitions, and by excluding those to the Γ-valley. The rate is plotted in Fig. **a**, and the phase space in Fig. **b**.

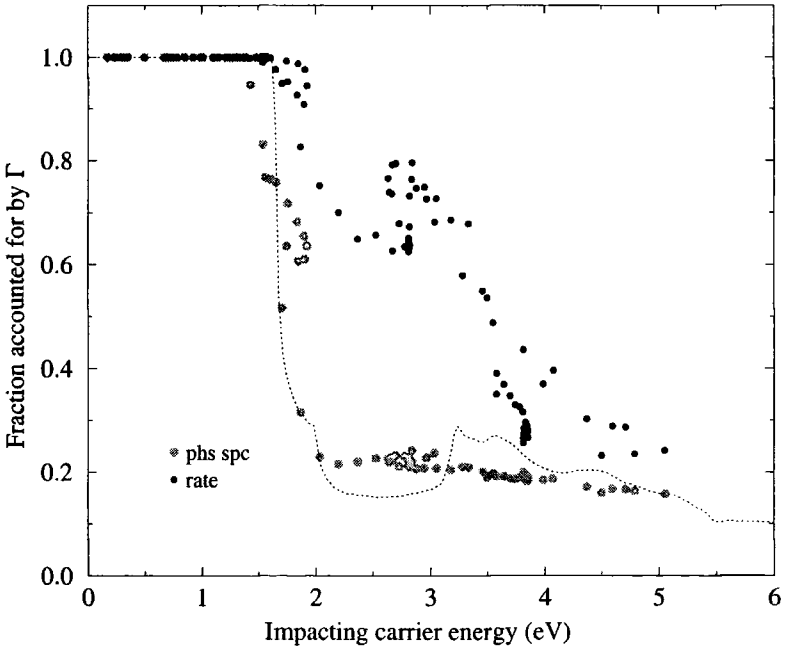


Figure 7.28: The electron initiated rate in InGaAs due to transitions involving at least one final state lying in the Γ-valley, expressed as a fraction of the total rate (the black points), and the equivalent data for the phase space (the grey points).

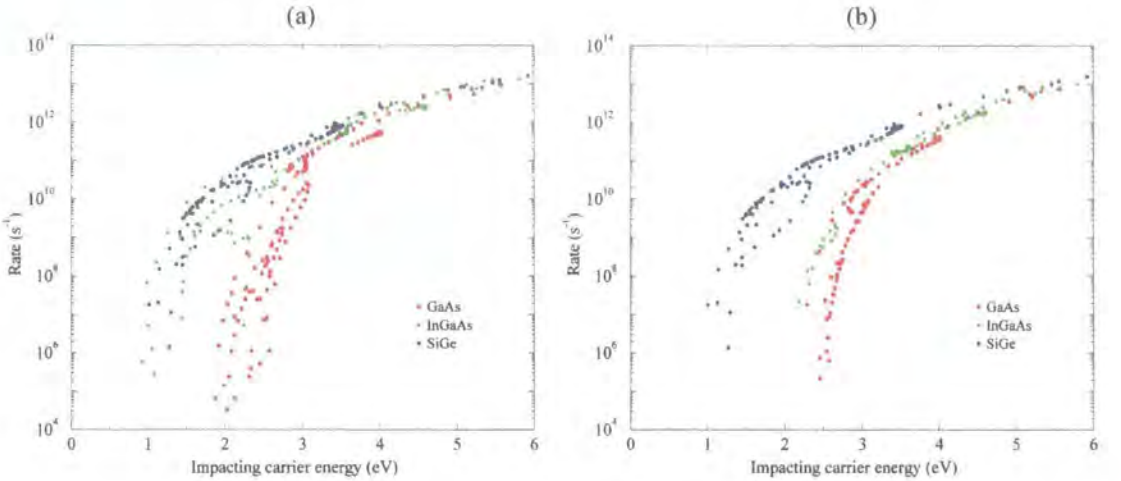


Figure 7.29: Electron initiated rates in GaAs, InGaAs and SiGe, obtained in the usual way by considering all transitions (plotted in Fig. **a**), and by excluding transitions to the Γ-valley (plotted in Fig. **b**).

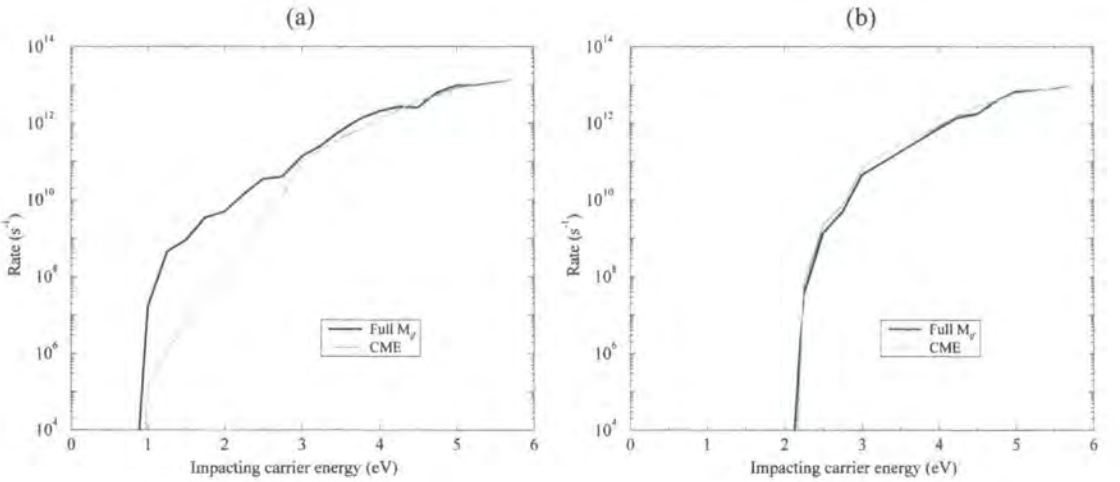


Figure 7.30: Electron initiated rates in InGaAs, calculated using the full expression for $|M_{if}|^2$, and using the CME approximation. In Fig. **a**, total rates are plotted. In Fig. **b**, rates calculated by excluding transitions to the Γ-valley.

7.4 Threshold Anisotropy and Softness of Rates

In this section the role of threshold anisotropy in affecting the softness of the rates is discussed. In order to investigate this, results presented in Chapter 6 will be re-examined.

In §6.3.2 of Chapter 6 the impact ionisation thresholds were plotted as a function of energy. It was shown that there does not generally exist a single energy above which impact ionisation can be initiated from any **k**-state, and below which it can be initiated from none, but rather a range of energies over which the fraction of ionising states increases from 0 to 1. This fraction is given by the expression (repeated from Eq. (6.2)):

$$f(E_i) = \frac{\int t(\mathbf{k})\ \delta(E(\mathbf{k}) - E_i)\ \mathrm{d}^3\mathbf{k}}{\int \delta(E(\mathbf{k}) - E_i)\ \mathrm{d}^3\mathbf{k}} \tag{7.6}$$

where $t(\mathbf{k})$ is defined as a function whose value is 1 if state **k** can initiate impact ionisation and zero otherwise. The range over which $0 < f(E_i) < 1$ was found to vary between materials and bands.

In §6.4.3, rates for each band in each material were presented as a function of energy. The rate plotted for a given energy is the mean rate due to all **k**-states at that energy, given by the expression (repeated from Eq. (6.3)):

$$R_{av}(E_i) = \frac{\int R(\mathbf{k})\ \delta(E(\mathbf{k}) - E_i)\ \mathrm{d}^3\mathbf{k}}{\int \delta(E(\mathbf{k}) - E_i)\ \mathrm{d}^3\mathbf{k}}. \tag{7.7}$$

It was noted in §6.4.3 that the rate averaged with respect to **k** at a given energy is an appropriate quantity to consider when carriers are scattered throughout the Brillouin zone, as they are when moving under the influence of a high field [17,25,54].

The integrals in Eq. (7.7) are performed for some particular value of E_i over all states, including those for which the rate is zero. We can obtain the mean rate at a particular energy due to only those states from which ionisation is possible from the

expression

$$R_{ion}(E_i) = \frac{\int t(\mathbf{k})R(\mathbf{k})\ \delta(E(\mathbf{k}) - E_i)\ \mathrm{d}^3\mathbf{k}}{\int t(\mathbf{k})\ \delta(E(\mathbf{k}) - E_i)\ \mathrm{d}^3\mathbf{k}}. \tag{7.8}$$

Using the result that $\int t(\mathbf{k})R(\mathbf{k})\delta(E) = \int R(\mathbf{k})\delta(E)$, Eqs. (7.6), (7.7) and (7.8) can be combined to give

$$R_{av}(E_i) = f(E_i) \times R_{ion}(E_i) \tag{7.9}$$

Since $f(E_i)$ is generally an increasing function of energy, we expect $R_{av}(E_i)$ to show softer threshold behaviour than $R_{ion}(E_i)$.

Sano *et al* [112] have studied the effect of anisotropy of the threshold on softness of the rate, concluding that rates in Si and GaAs are hard, but that greater anisotropy in the threshold of Si, combined with the fact that carriers are located at all points throughout the zone, leads to a softer effective rate. In terms of the notation used here, Sano *et al* claim that the dependence of R_{ion} on E_i is such as to give the same hard threshold behaviour in each material, and the variation in actual threshold softness seen in the values of R_{av} is due only to variation in $f(E_i)$. In view of this they suggest a **k**-vector dependent rate of the form

$$R(\mathbf{k}) = U\Big(E - E_0(\mathbf{k})\Big) \tag{7.10}$$

where $U(x)$ is the unit step function $U(x) = 0$ for $x < 0$ and $U(x) = 1$ for $x \geq 0$, and $E_0(\mathbf{k})$ is a **k**-dependent threshold energy. An altered form of this expression has been used in Monte Carlo simulations by Sano and co-workers to simulate Si [18,19] and by Chandramouli *et al* to simulate InP [113]. The expression they use is a variation of the Keldysh expression,

$$R(\mathbf{k}) = R_0\Big(E - E_0(\mathbf{k})\Big)^2 \tag{7.11}$$

Below, the results obtained in this work are analysed to determine if they show the same behaviour seen by Sano *et al.*

7.4.1 Comparison of Threshold Anisotropy

In Fig. 7.31, the values of $f(E_i)$ calculated here and by Sano *et al* are compared. The solid lines on the plot compare $f(E_i)$ calculated for electrons (in the first and second conduction bands) here and by Sano for GaAs. Agreement is quite good, and what differences there are can probably be attributed to differences in the band structure of each calculation (Sano uses the local pseudopotential band structure of Cohen and Bergstresser [89]). The dashed lines show $f(E_i)$ calculated here for SiGe and by Sano for Si. Since the lines are for different materials, they are not strictly comparable. However, Si and SiGe have similar band structure (an indirect gap with the bottom of the conduction band lying close to X), and we would expect to see qualitative similarity between the lines. It is surprising to see that the function of $f(E_i)$ obtained by Sano rises from 0 to 1 over a range of almost 5eV — considerably greater that the range of just 0.5eV obtained here for SiGe. In view of the fact that the largest range of any plot in §6.3.2 over which $0 < f(E_i) < 1$ is about 1.5eV (for the first conduction band of InGaAs), it seems highly unlikely that the threshold-finding algorithm of Beattie [61] used in this work would obtain such a large range had it been applied to Si. Sano has used the algorithm of Anderson and Crowell [106] to obtain the thresholds, which is known [20] to overestimate threshold anisotropy. We conclude therefore that the results for threshold anisotropy obtained here using Beattie's algorithm are to be preferred. The effect this has on the softness of the rate at threshold is examined below.

7.4.2 Effect of Anisotropy on Rates

Fig. 7.32 compares $R_{av}(E_i)$ (the overall mean rate) and $R_{ion}(E_i)$ (the mean rate due to only those states above threshold) calculated for electrons in SiGe. The energy range in which R_{av} is lower than R_{ion} corresponds to the energy range in which $0 < f(E_i) < 1$. The effect of the variation in $f(E_i)$ on the overall rate is not great. Table 7.3 gives fitting parameters for R_{av} and R_{ion} for carriers in each material. For R_{av}, the parameters A,

P and E_0 are all fitted as described in §6.4.3. For R_{ion}, A and P are fitted with E_0 fixed to the same value as obtained for R_{av}. From the values presented in the table it can be seen that when only those states able to initiate ionisation are considered, mean rates for both types of carrier show harder threshold behaviour, i.e. A increases and P decreases, as expected. However, the changes in A and P are not particularly great (in comparison to the differences seen between materials), confirming what can be seen from Fig. 7.32, i.e. that the effect of anisotropy in the thresholds plays only a small role in increasing the softness of the rates.

In the case of the electron initiated rates, it is interesting to note that for R_{ion} the P-parameter is the same for each material. However, fits to the mean volume of phase space as a function of energy do not give similar P-values for each material when the effect of anisotropy of the threshold is removed. Thus the equal P-values obtained for the rates is to some extent a coincidental combination of the different energy-dependencies of available phase space and matrix elements in each material and not too much significance should be read into this result. In addition, the softness of the threshold for R_{ion} still varies considerably between materials in that the fitted A-parameters vary (with GaAs showing the hardest threshold behaviour).

To summarise the information presented in this section, although the **k**-space anisotropy of the threshold has the effect of softening the threshold behaviour of the rates in each material studied, its influence is not great and the fact that the thresholds are soft (insofar as the fitted P parameters are large) is due mainly to the energy dependence of the rate itself on carrier energy rather than the dependence of the fraction of carriers at that energy that are above threshold. It seems likely that Sano *et al* have over estimated the anisotropy of the thresholds due to the use of Anderson and Crowell's threshold-finding algorithm.

		Fit to $R_{av}(E_i)$			Fit to $R_{ion}(E_i)$	
		A	P	E_0	A	P
e^-	GaAs	1.4×10^{11}	5.2	1.89	2.0×10^{11}	4.7
	InGaAs	1.6×10^{10}	5.6	0.75	4.2×10^{10}	4.7
	SiGe	4.6×10^{10}	4.9	0.84	6.1×10^{10}	4.7
h^+	GaAs	8.2×10^{10}	5.1	1.43	9.0×10^{10}	5.0
	InGaAs	1.5×10^{11}	4.2	0.73	1.6×10^{11}	4.2
	SiGe	7.8×10^{10}	4.7	1.23	1.1×10^{11}	4.5

Table 7.3: Fitting parameters for the mean rate R_{av} for all states at a particular energy calculated from Eq. (7.7), and for the mean rate R_{ion} due to only those states able to initiate impact ionisation calculated from Eq. (7.8).

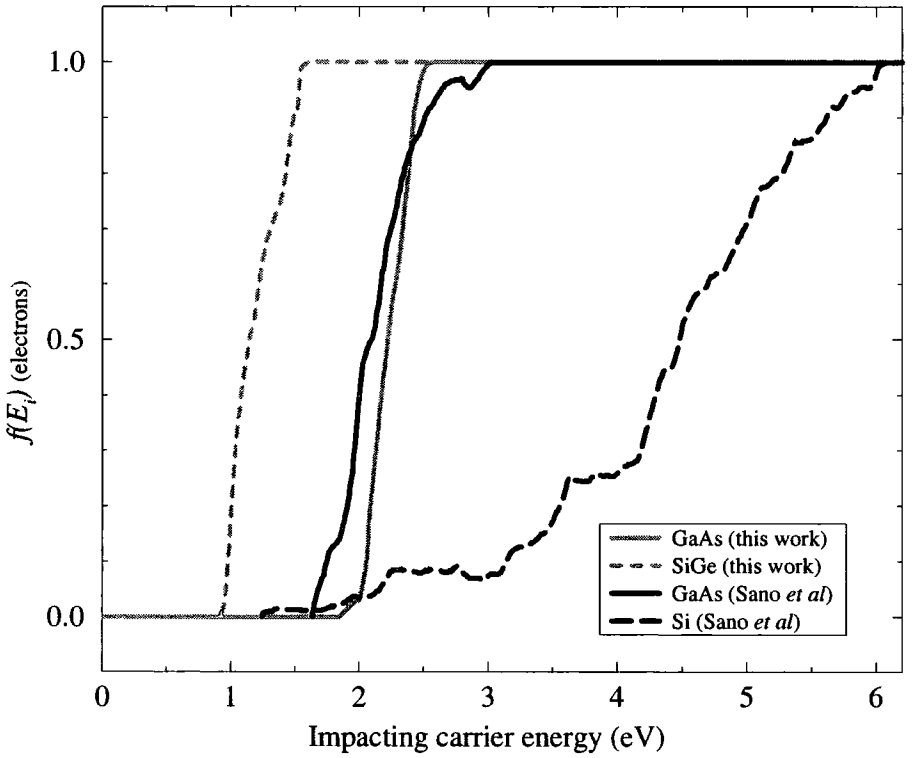


Figure 7.31: Comparison of threshold anisotropies obtained here for GaAs and SiGe and by Sano *et al* [111] for GaAs and Si. The function $f(E_i)$ is defined in Eq. (7.6)

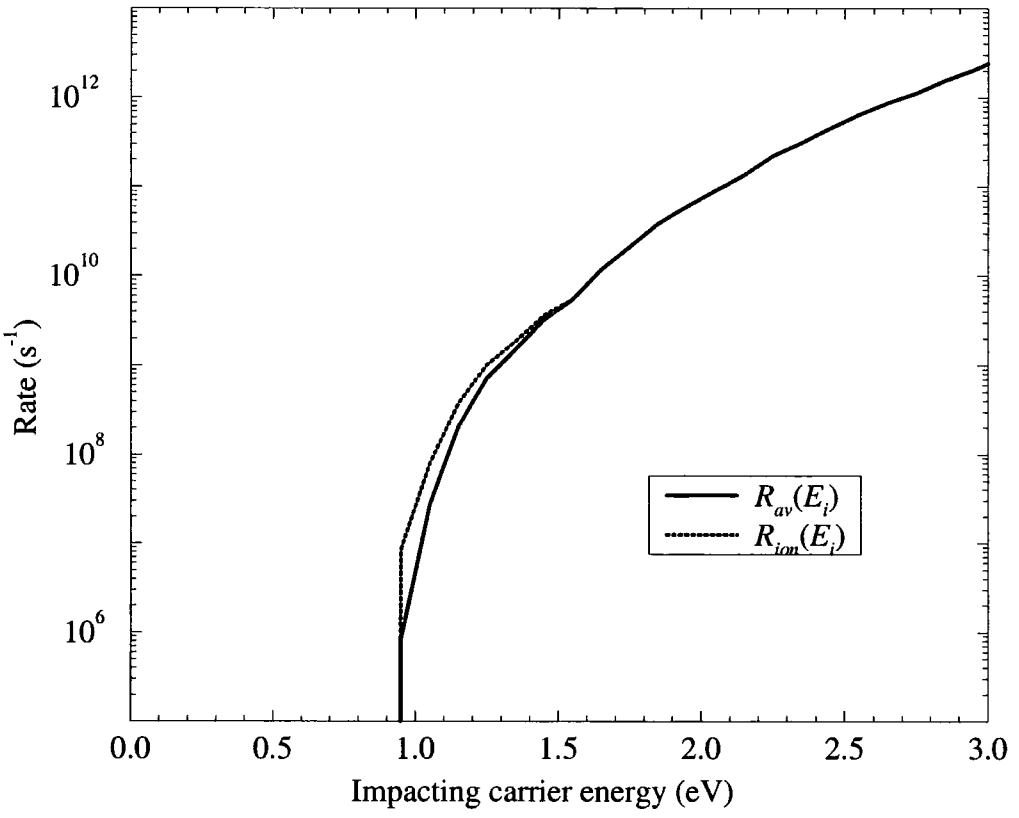


Figure 7.32: Mean rate R_{av} due to all states (Eq. (7.7)) compared to the mean rate R_{ion} due to only those states above threshold (Eq. (7.8)), plotted for SiGe.

Chapter 8

Conclusions

In this thesis, methods and results have been presented of band-to-band impact ionisation rate calculations carried out in the semi-classical Fermi's Golden Rule approximation for the materials GaAs, $In_{0.53}Ga_{0.47}As$ and $Si_{0.5}Ge_{0.5}$. The software developed to perform the calculations is in principle applicable to any unstrained diamond or zinc blende structure semiconductor.

Band structure for each material was obtained using the empirical pseudopotential method [81], discussed in Chapter 2. Pseudopotential parameters for InGaAs and SiGe were fitted by a Monte Carlo method (described in Chapter 3) to experimentally determined band gaps. Previously published pseudopotential parameters [81] are used for GaAs. 65 plane waves are used in the expansion of the pseudowavefunctions, which was found to give good convergence in calculated quantities such as the energy eigenvalues and dielectric function.

In order to rapidly obtain the band structure data at arbitrary **k**-vectors in the Brillouin zone, an interpolation scheme has been developed, as discussed in Chapter 3. Energies are interpolated quadratically from a mesh of pre-calculated points adapted to ensure that interpolation errors are kept uniformly low throughout the zone. Errors in energy values introduced by the interpolation scheme are typically less than a meV. The pseudowavefunctions are similarly quadratically interpolated from pre-calculated

expansion coefficients. More efficient use of computer memory is achieved by expanding wavefunctions at arbitrary **k**-vectors in terms of the wavefunctions at the zone centre. Uniform meshes of pre-calculated points are used to interpolate the expansion coefficients as it was found that, although adapted meshes were effective in reducing interpolation errors on the wavefunctions themselves, errors on matrix elements calculated from interpolated wavefunctions could not be effectively reduced by mesh adaptation. The errors introduced by interpolation on the wavefunctions themselves (with respect to wavefunctions obtained directly from the pseudopotential calculation) are a few percent ($\lesssim 4\%$). Errors on individual matrix elements incurred due to the interpolation of the wavefunctions were found to be considerably larger. However, quantities involving integration over many matrix elements, such as the impact ionisation rate, were found to be accurate to with a few percent.

The integration of the rate over all distinct energy and wavevector conserving transitions has been performed using two numerical algorithms, as described in Chapter 5. One is the surface integration algorithm of Beattie [61] and the other, which has been developed here, is a variation of Kane's algorithm [58] which (unlike Kane's) is efficient close to threshold. The rates obtained from each algorithm are in good agreement despite the quite different approaches employed by each, indicating that they are numerically reliable.

The calculation of the impact ionisation transition matrix elements was discussed in Chapter 4. The matrix elements are calculated using the pseudowavefunctions obtained from the pseudopotential calculation, via the interpolation scheme, and include the terms which are commonly neglected in calculations for narrow band gap semiconductors [82]. The **q**- and ω-dependent expression for the dielectric function was calculated from the pseudopotential band structure using the expression given by Walter and Cohen [83]. An isotropic approximation to this function (i.e. $\epsilon(q,\omega) \simeq \epsilon(\mathbf{q},\omega)$) is used in the evaluation of the matrix elements. The error incurred in the calculated rate due to the use of this isotropic approximation is estimated to be less than $\sim 5\%$.

The convergence of the matrix elements with respect to the number of plane waves used to expand the pseudowavefunctions was tested. In InGaAs it was found to be good (to within a few percent) and it was assumed similarly good convergence would be obtained in GaAs. In SiGe the convergence was found to be considerably worse ($\sim$ 30%), due to the fact that in this material transitions generally involve greater **q**-transfer which does not favour rapid convergence of the matrix elements.

The aspects of the rate calculation described above were combined to obtain impact ionisation rates for electrons and holes in each of the three materials studied. In common with many rate calculations using real band structure [10,22,25,26,65,66], the rates were found to be explicitly dependent on the **k**-vector of the impacting carrier, and not on just its energy as is the case in the Keldysh formula [53]. This is due to the restrictive nature of the requirement for simultaneous energy and momentum conservation. Thus carriers at the same energy but different positions in **k**-space can have widely varying rates. These explicitly **k**-dependent rates were approximated by a function of energy alone of the form (repeated from Eq. (6.4) of Chapter 6)

$$R(E) = A(E - E_0)^P, \tag{8.1}$$

where E is the impacting carrier energy and A, P and E_0 are fitted parameters. Values of A, P and E_0 for each material are listed in Table 6.7 of Chapter 6. It was generally found that the best fit was obtained with $P > 2$, in common with other realistic band structure calculations presented in the literature [26,28,67]. The Keldysh formula [53] for the rate is of the form of Eq (8.1) with $P = 2$, which is obtained assuming a direct gap, spherical parabolic band structure and constant matrix elements. The greater value of P obtained here is due to the deviation of the real band structure from the idealised parabolic case, and indicates a softer threshold.

Rates obtained in this work were compared with those obtained by several other workers [20–22,26,27,59,60,105,111]. Reasonably good agreement was found in GaAs, but in InGaAs and SiGe rates obtained here and by other authors varied considerably. The

effect on the rates of using different band structure and different approximations in evaluating the matrix elements was investigated and found to be relatively small in comparison to discrepancies between authors, particularly for InGaAs and SiGe, and so it was concluded that the details of the implementation of the numerical rate integration accounted for much of the variation. It was noted, however, that the effects of differences in the band structure was not fully explored, and that in materials for which limited experimental data is available, variation in band structure may have a greater influence than was determined here.

The impact ionisation thresholds were found using the algorithm of Beattie [107], rather than the commonly used algorithm of Anderson and Crowell [106] which is known to give inaccurate estimates for the thresholds under certain conditions [20]. The thresholds were found to be highly anisotropic in **k**-space, particularly in the case of electron initiated processes, reflecting the anisotropy of the energy bands themselves. Tests of whether this **k**-space anisotropy leads to anisotropy in the α-coefficient are typically carried out by applying fields in the 100, 110 and 111 directions [17,41–43]. However, in GaAs it was found that the shape of the thresholds in **k**-space was such that ballistic electrons would be most likely to reach threshold when travelling in the 210-direction, and therefore any anisotropy in the α-coefficient, if it exists, would be most clearly seen for fields oriented along this axis.

It was found that the threshold cannot be characterised in terms of a single energy above which impact ionisation can be initiated from any **k**-state and below which it can be initiated from none. Instead, in each material the fraction of all points in **k**-space at a particular energy from which carriers can initiate impact ionisation was found to increase from 0 to 1 over an energy range typically of the order of 1 eV. It was found that this gradual rise in the number of ionising states with respect to energy had the effect of softening the threshold, as predicted by Sano *et al* [111]. However the degree of softening introduced by the threshold energy range was found to be much smaller than that suggested by Sano, the softness obtained from the rate calculations being

mainly due to the gradual rise in the volume of available phase space of final states as the threshold energy is exceeded rather than the rise in the number of ionising states at a given energy. The apparent over-estimation by Sano *et al* of the importance of the threshold anisotropy is probably due to their use of Anderson and Crowell's [106] threshold-finding algorithm, which is known to give inaccurate results under certain circumstances [20].

The distribution throughout **k**-space of the secondary states (i.e. the impacted and final states) has been examined. Generated carriers were found to be confined to the conduction band valleys and the top of the valence band for low energy impacting carriers, whereas they are distributed throughout the zone for the highest energy impacting carriers. In the direct gap materials studied (GaAs and InGaAs), the distribution of generated carriers was found to vary significantly depending on the position in **k**-space of the impacting carrier. In the indirect gap material (SiGe) the distributions of generated carriers were found to be similar for different impacting carriers. In all materials, the mean energy of the generated carriers was found to be approximately proportional to the energy of the impacting carriers. In the direct gap materials, the generated electrons on average each take a slightly greater share than the generated holes of the kinetic energy made available by the impacting carrier (whether it is an electron or a hole), while in the indirect gap material, the opposite was found.

The individual contributions of the volume of available phase space and the matrix elements to the overall rate was examined. The rate and the volume of phase space were found to be in good quantitative agreement in SiGe (to within a scaling factor corresponding to the mean matrix element), while in the direct gap materials, although there was a qualitative correspondence between the two, poor quantitative agreement was observed. In the direct gap materials, particularly InGaAs, the P-parameter of Eq. (8.1) was found to be lower when fitted to the rate than the volume of phase space as a function of energy, i.e. the threshold behaviour of the rate was harder than that of the phase space. In SiGe, the fits for the rate and volume of phase space were very

similar. It was concluded that approximation of the matrix elements by a constant expression will lead to a softening of predicted electron and hole initiated rates in direct gap materials, but give a good estimate of the electron initiated rate in indirect gap materials. General predictions regarding the softness of hole initiated transitions for indirect gap materials could not be made due to the complicated influence on the threshold softness of the closely spaced thresholds for the light, heavy and spin split off bands.

The role of the matrix elements in influencing the distribution of secondary states was examined. For electron initiated rates in the direct gap materials it was found that the matrix elements act to enhance the low **q**-transfer transitions, particularly for impacting carriers in the second conduction band. For electron initiated transitions in SiGe and hole initiated transitions in all the materials, the matrix elements were found to have little effect on the secondary state distribution.

It has been noted throughout this thesis that the properties of the direct gap materials studied are frequently in qualitative agreement, while those of the indirect gap material differ. In the direct gap materials the role of the Γ-valley (which is only very shallow in SiGe) in influencing the rates was found to be considerably greater than would be expected from its density of states in comparison to the higher effective mass satellite valleys. This in turn was found to be due to the small **q**-transfer associated with transitions involving this valley, and hence enhancement of the corresponding matrix elements. The contribution to the rate of the Γ-valley has the effect of softening the threshold, particularly in InGaAs where the Γ-satellite separation is a considerably greater fraction of the band gap than in GaAs.

Suggestions for Further Work

The calculations in this thesis were performed in the semi-classical Fermi's Golden Rule approximation, but high fields and high phonon scattering rates at the energies at which impact ionisation is typically of interest reduce the applicability of this ap-

proximation. An important extension of this work would be to incorporate a fuller treatment of these effects into the rate calculation. Both the intracollisional effect (applicable at high fields) and collision broadening (applicable at high scattering rates) have the effect of relaxing the requirement for energy conservation, leading to a softening of the threshold behaviour of the rates and an increase in anisotropy [69–72]. In fact, under conditions of non-energy conservation, the concept of a threshold becomes inapplicable, and ionisation can be initiated by carriers well below the semi-classical minimum energy.

The number of materials studied here is limited, due to considerations of available time and computer resources. To fully explore the differences noted between the direct and indirect materials studied here, and to fully investigate other trends in material properties, other semiconductors must be considered. In particular, Allam [67] has argued that the semiconductors Si (which has similar band structure to $Si_{0.5}Ge_{0.5}$), GaAs, InAs and $In_{0.53}Ga_{0.47}As$ are similar in that they have similar values of $\langle E_{ind} \rangle$ (defined in Eq. (7.5) of Chapter 7). Different behaviour may be seen in InP for example, which has a higher value of $\langle E_{ind} \rangle$. It would also be desirable to include the effects of strain in the calculations, and to apply them to semiconductors of the wurtzite structure such as GaN.

Ultimately, to gain most insight into the role of impact ionisation in devices, the results of the rate calculations performed here must be incorporated into a full band transport simulation, for which the Monte Carlo method is a suitable technique. Unfortunately, the computational effort required for numerical modelling of this sort is very great, both in terms of the development and running of the software.

Appendix A

Wavefunctions and Basis Sets

An electron in band b with wave vector $\mathbf{k}$, is described by the wavefunction

$$\psi_b(\mathbf{r},\mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_b(\mathbf{r},\mathbf{k}) \tag{A.1}$$

where $u_b(\mathbf{r},\mathbf{k})$ is the Bloch periodic part, and is expressed in terms of an expansion of plane waves as

$$u_b(\mathbf{r},\mathbf{k}) = \frac{1}{\sqrt{\Omega}} \sum_i \alpha_{b,i}(\mathbf{k}) e^{i\mathbf{G}_i\cdot\mathbf{r}} \tag{A.2}$$

where Ω is the crystal volume and the coefficient $\alpha_{b,i}$ is in general a complex number.

The Bloch periodic part $u_b(\mathbf{r},\mathbf{k})$ may also be expressed as an expansion in terms of another basis set — the zone centre wavefunctions — consisting of orthonormal-normal functions $\phi_j(\mathbf{r})$:

$$u_b(\mathbf{r},\mathbf{k}) = \sum_j \beta_{b,j}(\mathbf{k}) \phi_j(\mathbf{r}) \tag{A.3}$$

where $\phi_j(\mathbf{r})$ is the wavefunction evaluated at $\mathbf{k} = 0$ for the j^{th} band, and the coefficient $\beta_{b,j}$ is a complex number.

If the zone centre wavefunctions $\phi_j(\mathbf{r})$ are themselves expressed as an expansion in

terms of plane waves:

$$\phi_j(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} \sum_k \gamma_{j,k} e^{i\mathbf{G}_k \cdot \mathbf{r}} \tag{A.4}$$

where again $\gamma_{j,k}$ is a complex coefficient, then we can convert $u_b(\mathbf{r}, \mathbf{k})$ from an expansion in terms of plane waves, as in Eq. (A.2), to an expansion in terms of zone centre wavefunctions, as in Eq. (A.3), and vice-versa.

A.1 Plane Wave to Zone Centre Conversion

Suppose we have a Bloch function expanded as in Eq. (A.2) and we would like it expanded as in Eq. (A.3). In other words, we know the set of coefficients $\alpha_{b,i}$ and would like to know the set $\beta_{b,j}$.

Multiplying both sides of Eq. (A.3) with $\phi_p^*(\mathbf{r})$ and integrating with respect to $\mathbf{r}$ over the volume of the crystal, we get

$$\int_\Omega \phi_p^*(\mathbf{r})\ u_b(\mathbf{r}, \mathbf{k})\ d\mathbf{r} = \beta_{b,p}(\mathbf{k}). \tag{A.5}$$

Using Eq. (A.2) and Eq. (A.4) to replace the expressions on the left-hand-side of Eq. (A.5), we get

$$\frac{1}{\Omega} \int_\Omega \left[\sum_k \gamma_{p,k}^* e^{-i\mathbf{G}_k \cdot \mathbf{r}} \right] \left[\sum_i \alpha_{b,i}(\mathbf{k}) e^{i\mathbf{G}_i \cdot \mathbf{r}} \right] d\mathbf{r} = \beta_{b,p}(\mathbf{k}). \tag{A.6}$$

Doing the integral gives

$$\boxed{\beta_{b,p}(\mathbf{k}) = \sum_i \alpha_{b,i}(\mathbf{k})\ \gamma_i^*(p)} \tag{A.7}$$

Thus, to convert a plane wave expansion of the wavefunction for an electron at $\mathbf{k}$ in band b to a zone centre expansion, it is necessary to perform the summation Eq. (A.7) for each of the zone centre coefficients $\beta_{b,p}$ in the expansion.

A.2 Zone Centre to Plane Wave Conversion

The opposite process to that discussed in §A.1 is to convert a zone centre expansion to a plane wave expansion — that is, to convert the set of coefficients $\beta_{b,j}$ to the set $\alpha_{b,i}$.

Using Eq. (A.4) to replace the $\phi_j(\mathbf{r})$ in Eq. (A.3), we get

$$\psi_b(\mathbf{r},\mathbf{k}) = \sum_j \beta_{b,j}(\mathbf{k}) \frac{1}{\sqrt{\Omega}} \sum_k \gamma_{j,k} e^{i\mathbf{G}_k\cdot\mathbf{r}} \tag{A.8}$$

which can be re-arranged to give

$$\psi_b(\mathbf{r},\mathbf{k}) = \frac{1}{\sqrt{\Omega}} \sum_k \left[\sum_j \beta_{b,j}(\mathbf{k})\gamma_{j,k}\right] e^{i\mathbf{G}_k\cdot\mathbf{r}}. \tag{A.9}$$

Comparing the term in square brackets with $\alpha_{b,i}(\mathbf{k})$ in Eq. (A.2) we can see that

$$\boxed{\alpha_{b,p}(\mathbf{k}) = \sum_j \beta_{b,j}(\mathbf{k})\ \gamma_{j,p}} \tag{A.10}$$

A.3 The Zone Centre Wavefunctions Themselves

Conversion between the zone centre and plane wave representations of the wavefunction, using the boxed equations Eq. (A.7) and Eq. (A.10), requires a knowledge of the set of coefficients $\gamma_{j,k}$ used to expand the zone centre wavefunctions in Eq. (A.4).

The pseudopotential method is used to generate the wavefunctions at the zone centre, in just the same way as it is used at general **k**-points throughout the Brillouin zone. However, at the Γ-point, all the bands are at least doubly-degenerate and so the corresponding wavefunctions will be output in an arbitrary linear combination which in general is not symmetrised as in Figs. 3.12 and 3.13, for example. The symmetry operations listed in Table 3.2 are most easily applied to symmetrised wavefunctions, and so the $\gamma_{j,k}$ obtained for sets of degenerate bands are linearly recombined so as to ensure that the new zone centre wavefunctions have the required symmetry. When this is done, the set of $\gamma_{j,k}$ is stored for use in performing the conversion between basis sets.

Appendix B

Matrix Element with Spin and Exchange

In this appendix, the expression for the matrix element summation including spin and exchange terms is set out in detail.

The effect of including spin in the calculation of the direct matrix element M_d is discussed in §4.2.5. The expression for the full matrix element M_{if}, including direct and exchange parts, is given by

$$M_{if} = M_d - M_e \tag{B.1}$$

where

$$\begin{aligned} M_d = \int \Big[& {}^{\uparrow}\psi^*_{1'}(\mathbf{r}_1){}^{\uparrow}\psi^*_{2'}(\mathbf{r}_2)V^{\uparrow}\psi^*_1(\mathbf{r}_1){}^{\uparrow}\psi^*_2(\mathbf{r}_2) \\ + & {}^{\uparrow}\psi^*_{1'}(\mathbf{r}_1){}^{\downarrow}\psi^*_{2'}(\mathbf{r}_2)V^{\uparrow}\psi^*_1(\mathbf{r}_1){}^{\downarrow}\psi^*_2(\mathbf{r}_2) \\ + & {}^{\downarrow}\psi^*_{1'}(\mathbf{r}_1){}^{\uparrow}\psi^*_{2'}(\mathbf{r}_2)V^{\downarrow}\psi^*_1(\mathbf{r}_1){}^{\uparrow}\psi^*_2(\mathbf{r}_2) \\ + & {}^{\downarrow}\psi^*_{1'}(\mathbf{r}_1){}^{\downarrow}\psi^*_{2'}(\mathbf{r}_2)V^{\downarrow}\psi^*_1(\mathbf{r}_1){}^{\downarrow}\psi^*_2(\mathbf{r}_2) \Big] \, d^3\mathbf{r}_1 d^3\mathbf{r}_2 \end{aligned} \tag{B.2}$$

and

$$
\begin{aligned}
M_e = \int \Big[& \; {}^{\uparrow}\psi^*_{2'}(\mathbf{r}_1){}^{\uparrow}\psi^*_{1'}(\mathbf{r}_2) V {}^{\uparrow}\psi^*_{1}(\mathbf{r}_1){}^{\uparrow}\psi^*_{2}(\mathbf{r}_2) \\
& + {}^{\uparrow}\psi^*_{2'}(\mathbf{r}_1){}^{\downarrow}\psi^*_{1'}(\mathbf{r}_2) V {}^{\uparrow}\psi^*_{1}(\mathbf{r}_1){}^{\downarrow}\psi^*_{2}(\mathbf{r}_2) \\
& + {}^{\downarrow}\psi^*_{2'}(\mathbf{r}_1){}^{\uparrow}\psi^*_{1'}(\mathbf{r}_2) V {}^{\downarrow}\psi^*_{1}(\mathbf{r}_1){}^{\uparrow}\psi^*_{2}(\mathbf{r}_2) \\
& + {}^{\downarrow}\psi^*_{2'}(\mathbf{r}_1){}^{\downarrow}\psi^*_{1'}(\mathbf{r}_2) V {}^{\downarrow}\psi^*_{1}(\mathbf{r}_1){}^{\downarrow}\psi^*_{2}(\mathbf{r}_2) \Big] \, d^3\mathbf{r}_1 d^3\mathbf{r}_2
\end{aligned}
\tag{B.3}
$$

Eq. (4.16) gives an expression for the direct part of the matrix element, calculated without spin. Here the expression is generalised to included spin and exchange.

Each wavefunction is a linear combination of spin-up and spin-down parts, with each of these parts being represented as a sum of plane waves:

$$
\psi_\alpha = \frac{1}{\sqrt{\Omega}} \Big[\, |\uparrow\rangle \sum_{\mathbf{G}_\alpha} {}^{\uparrow}A_\alpha(\mathbf{G}_\alpha) e^{i(\mathbf{k}_\alpha + \mathbf{G}_\alpha)\cdot\mathbf{r}} \; + \; |\downarrow\rangle \sum_{\mathbf{G}_\alpha} {}^{\downarrow}A_\alpha(\mathbf{G}_\alpha) e^{i(\mathbf{k}_\alpha + \mathbf{G}_\alpha)\cdot\mathbf{r}} \, \Big]. \tag{B.4}
$$

This form of the wavefunction is substituted into Eqs. (B.2) and (B.3). Using the expression for V given by Eq. (4.12) and the result of Eq. (4.15), we get

$$
\begin{aligned}
M_d = \sum_{\mathbf{G}_1,\mathbf{G}_2,\mathbf{G}_{1'},\mathbf{G}_{2'}} & \frac{e^2 \, \delta_{\mathbf{G}_1+\mathbf{G}_2-\mathbf{G}_{1'}-\mathbf{G}_{2'}+\mathbf{k}_1+\mathbf{k}_2-\mathbf{k}_{1'}-\mathbf{k}_{2'}}}{\Omega \, \epsilon_0 \epsilon(\mathbf{q}_d,\omega_d) \, |\mathbf{q}_d|^2} \times \\
& \Big[\; {}^{\uparrow}A^*_{1'}(\mathbf{G}_{1'}) \, {}^{\uparrow}A^*_{2'}(\mathbf{G}_{2'}) \, {}^{\uparrow}A_1(\mathbf{G}_1) \, {}^{\uparrow}A_2(\mathbf{G}_2) \\
& + {}^{\uparrow}A^*_{1'}(\mathbf{G}_{1'}) \, {}^{\downarrow}A^*_{2'}(\mathbf{G}_{2'}) \, {}^{\uparrow}A_1(\mathbf{G}_1) \, {}^{\downarrow}A_2(\mathbf{G}_2) \\
& + {}^{\downarrow}A^*_{1'}(\mathbf{G}_{1'}) \, {}^{\uparrow}A^*_{2'}(\mathbf{G}_{2'}) \, {}^{\downarrow}A_1(\mathbf{G}_1) \, {}^{\uparrow}A_2(\mathbf{G}_2) \\
& + {}^{\downarrow}A^*_{1'}(\mathbf{G}_{1'}) \, {}^{\downarrow}A^*_{2'}(\mathbf{G}_{2'}) \, {}^{\downarrow}A_1(\mathbf{G}_1) \, {}^{\downarrow}A_2(\mathbf{G}_2) \Big]
\end{aligned}
\tag{B.5}
$$

where

$$
\mathbf{q}_d = \mathbf{G}_1 - \mathbf{G}_{1'} + \mathbf{k}_1 - \mathbf{k}_{1'} \tag{B.6}
$$

$$
\hbar\omega_d = E(\mathbf{k}_1) - E(\mathbf{k}_{1'}) \tag{B.7}
$$

and

$$M_e = \sum_{\mathbf{G}_1,\mathbf{G}_2,\mathbf{G}_{1'},\mathbf{G}_{2'}} \frac{e^2\ \delta_{\mathbf{G}_1+\mathbf{G}_2-\mathbf{G}_{1'}-\mathbf{G}_{2'}+\mathbf{k}_1+\mathbf{k}_2-\mathbf{k}_{1'}-\mathbf{k}_{2'}}}{\Omega\ \epsilon_0\epsilon(\mathbf{q}_e,\omega_e)\ |\mathbf{q}_e|^2} \times$$

$$\begin{aligned}
&\Big[\ {}^{\uparrow}\mathcal{A}^*_{2'}(\mathbf{G}_{1'})\ {}^{\uparrow}\mathcal{A}^*_{1'}(\mathbf{G}_{2'})\ {}^{\uparrow}\mathcal{A}_1(\mathbf{G}_1)\ {}^{\uparrow}\mathcal{A}_2(\mathbf{G}_2) \\
&+\ {}^{\uparrow}\mathcal{A}^*_{2'}(\mathbf{G}_{1'})\ {}^{\downarrow}\mathcal{A}^*_{1'}(\mathbf{G}_{2'})\ {}^{\uparrow}\mathcal{A}_1(\mathbf{G}_1)\ {}^{\downarrow}\mathcal{A}_2(\mathbf{G}_2) \\
&+\ {}^{\downarrow}\mathcal{A}^*_{2'}(\mathbf{G}_{1'})\ {}^{\uparrow}\mathcal{A}^*_{1'}(\mathbf{G}_{2'})\ {}^{\downarrow}\mathcal{A}_1(\mathbf{G}_1)\ {}^{\uparrow}\mathcal{A}_2(\mathbf{G}_2) \\
&+\ {}^{\downarrow}\mathcal{A}^*_{2'}(\mathbf{G}_{1'})\ {}^{\downarrow}\mathcal{A}^*_{1'}(\mathbf{G}_{2'})\ {}^{\downarrow}\mathcal{A}_1(\mathbf{G}_1)\ {}^{\downarrow}\mathcal{A}_2(\mathbf{G}_2)\ \Big]
\end{aligned} \tag{B.8}$$

where

$$\mathbf{q}_e = \mathbf{G}_1 - \mathbf{G}_{1'} + \mathbf{k}_1 - \mathbf{k}_{2'} \tag{B.9}$$

$$\hbar\omega_e = E(\mathbf{k}_1) - E(\mathbf{k}_{2'}) \tag{B.10}$$

The factorisation of the direct matrix element without spin is discussed in §4.2.4. Here the factorisation of the general expression for the matrix element, including exchange and spin terms, is given.

If the matrix element is written

$$M_{if} = M_d - M_e = \frac{e^2}{\epsilon_0\Omega}S = \frac{e^2}{\epsilon_0\Omega}(S_d - S_e) \tag{B.11}$$

then S_d and S_e, which can be obtained by factorisation of Eq. (B.5) and Eq. (B.8), are

given by

$$
\begin{aligned}
S_d = \sum_{\mathbf{G}_\Delta} \Bigg\{ & \Big[\sum_{\mathbf{G}_1} {}^{\uparrow}\mathcal{A}^*_{1'}(\mathbf{G}_1 - \mathbf{G}_\Delta)\ {}^{\uparrow}\mathcal{A}_1(\mathbf{G}_1)\Big]\Big[\sum_{\mathbf{G}_{2'}} {}^{\uparrow}\mathcal{A}^*_{2'}(\mathbf{G}_{2'})\ {}^{\uparrow}\mathcal{A}_2(\mathbf{G}_{2'} - \mathbf{G}_\Delta + \mathbf{G}_u)\Big] \\
+ & \Big[\sum_{\mathbf{G}_1} {}^{\uparrow}\mathcal{A}^*_{1'}(\mathbf{G}_1 - \mathbf{G}_\Delta)\ {}^{\uparrow}\mathcal{A}_1(\mathbf{G}_1)\Big]\Big[\sum_{\mathbf{G}_{2'}} {}^{\downarrow}\mathcal{A}^*_{2'}(\mathbf{G}_{2'})\ {}^{\downarrow}\mathcal{A}_2(\mathbf{G}_{2'} - \mathbf{G}_\Delta + \mathbf{G}_u)\Big] \\
+ & \Big[\sum_{\mathbf{G}_1} {}^{\downarrow}\mathcal{A}^*_{1'}(\mathbf{G}_1 - \mathbf{G}_\Delta)\ {}^{\downarrow}\mathcal{A}_1(\mathbf{G}_1)\Big]\Big[\sum_{\mathbf{G}_{2'}} {}^{\uparrow}\mathcal{A}^*_{2'}(\mathbf{G}_{2'})\ {}^{\uparrow}\mathcal{A}_2(\mathbf{G}_{2'} - \mathbf{G}_\Delta + \mathbf{G}_u)\Big] \\
+ & \Big[\sum_{\mathbf{G}_1} {}^{\downarrow}\mathcal{A}^*_{1'}(\mathbf{G}_1 - \mathbf{G}_\Delta)\ {}^{\downarrow}\mathcal{A}_1(\mathbf{G}_1)\Big]\Big[\sum_{\mathbf{G}_{2'}} {}^{\downarrow}\mathcal{A}^*_{2'}(\mathbf{G}_{2'})\ {}^{\downarrow}\mathcal{A}_2(\mathbf{G}_{2'} - \mathbf{G}_\Delta + \mathbf{G}_u)\Big] \\
& \Bigg\} \times \frac{1}{\epsilon(\mathbf{q}_d, \omega_d)|\mathbf{q}_d|^2}
\end{aligned}
\tag{B.12}
$$

and

$$
\begin{aligned}
S_e = \sum_{\mathbf{G}_\Delta} \Bigg\{ & \Big[\sum_{\mathbf{G}_1} {}^{\uparrow}\mathcal{A}^*_{2'}(\mathbf{G}_1 - \mathbf{G}_\Delta)\ {}^{\uparrow}\mathcal{A}_1(\mathbf{G}_1)\Big]\Big[\sum_{\mathbf{G}_{2'}} {}^{\uparrow}\mathcal{A}^*_{1'}(\mathbf{G}_{2'})\ {}^{\uparrow}\mathcal{A}_2(\mathbf{G}_{2'} - \mathbf{G}_\Delta + \mathbf{G}_u)\Big] \\
+ & \Big[\sum_{\mathbf{G}_1} {}^{\uparrow}\mathcal{A}^*_{2'}(\mathbf{G}_1 - \mathbf{G}_\Delta)\ {}^{\uparrow}\mathcal{A}_1(\mathbf{G}_1)\Big]\Big[\sum_{\mathbf{G}_{2'}} {}^{\downarrow}\mathcal{A}^*_{1'}(\mathbf{G}_{2'})\ {}^{\downarrow}\mathcal{A}_2(\mathbf{G}_{2'} - \mathbf{G}_\Delta + \mathbf{G}_u)\Big] \\
+ & \Big[\sum_{\mathbf{G}_1} {}^{\downarrow}\mathcal{A}^*_{2'}(\mathbf{G}_1 - \mathbf{G}_\Delta)\ {}^{\downarrow}\mathcal{A}_1(\mathbf{G}_1)\Big]\Big[\sum_{\mathbf{G}_{2'}} {}^{\uparrow}\mathcal{A}^*_{1'}(\mathbf{G}_{2'})\ {}^{\uparrow}\mathcal{A}_2(\mathbf{G}_{2'} - \mathbf{G}_\Delta + \mathbf{G}_u)\Big] \\
+ & \Big[\sum_{\mathbf{G}_1} {}^{\downarrow}\mathcal{A}^*_{2'}(\mathbf{G}_1 - \mathbf{G}_\Delta)\ {}^{\downarrow}\mathcal{A}_1(\mathbf{G}_1)\Big]\Big[\sum_{\mathbf{G}_{2'}} {}^{\downarrow}\mathcal{A}^*_{1'}(\mathbf{G}_{2'})\ {}^{\downarrow}\mathcal{A}_2(\mathbf{G}_{2'} - \mathbf{G}_\Delta + \mathbf{G}_u)\Big] \\
& \Bigg\} \times \frac{1}{\epsilon(\mathbf{q}_e, \omega_e)|\mathbf{q}_e|^2}
\end{aligned}
\tag{B.13}
$$

Bibliography

[1] V. S. Vavilov, Phys. Chem. Solids **8**, 223 (1959)

[2] S. M. Sze, *Semiconductor Devices, Physics and Technology*, chapter 3, Wiley (1985)

[3] C. M. Hu, S. C. Tam, F. C. Hsu, P. K. Ko, T. Y. Chan and K. W. Terril, IEEE Trans. Electron Devices **32**, 375 (1985)

[4] K. Hui, C. Hu, P. George and P. K. Ko, IEEE Electron Device Lett. **11**, 113 (1990)

[5] C. Canali, A. Paccagnella, P. Pisoni, C. Tedesco, P. Telaroli and E. Zanoli, IEEE Trans. Electron Devices **38**, 2571 (1991)

[6] M. H. Somerville, J. A. del Alamo and W. Hoke, IEEE Electron Device Lett. **17**, 473 (1996)

[7] I. M. Hafez, G. Ghibaudo and F. Balestra, IEEE Trans. Electron Devices. **37**, 818 (1990)

[8] B. Georgescu, M. A. Py, A. Souifi, G. Post and G. Guillot, IEEE Electron Device Lett. **19**, 154 (1998)

[9] J. Haruyama and H. Katano, J. Appl. Phys. **77**, 3913 (1995)

[10] H. Mizuno, M. Morifuji, K. Taniguchi and C. Hamaguchi, J. Appl. Phys. **74**, 1100 (1993)

[11] T. H. Ning, P. W. Cook, R. H. Dennard, C. M. Osburn, S. E. Schuster and H. Yu, IEEE J. Solid State Circuits **14**, 268 (1979)

[12] K. Fucuda, H. J. Peifer, B. Meinerzhagen, R. Thoma and W. L. Engl, Jpn. J. Appl. Phys. Part 1 **31**, 3763 (1992)

[13] G. E. Stillman and C. M. Wolf, *Infrared Detectors II*, volume 12 of *Semiconductors and Semimetals*, chapter 5, Academic (1977)

[14] F. Capasso, *Lightwave Communications Technology*, volume 22 of *Semiconductors and Semimetals*, chapter 1, Academic (1985)

[15] S. M. Sze, *Physics of Semiconductor Devices*, Wiley, 2nd edition (1981)

[16] J. D. Bude, IEEE Electron Device Lett. **16**, 439 (1995)

[17] H. Scichijo and K. Hess, Phys. Rev. B **23**, 4197 (1981)

[18] N. Sano, M. Tomizawa and A. Yoshii, Appl. Phys. Lett. **56**, 653 (1990)

[19] N. Sano, T. Aoki, M. Tomizawa and A. Yoshii, Phys. Rev. B **41**, 12122 (1990)

[20] J. Bude and K. Hess, J. Appl. Phys. **72**, 3554 (1992)

[21] Y. Wang and K. F. Brennan, J. Appl. Phys. **76**, 974 (1994)

[22] Y. Kamakura, H. Mizuno, M. Yamaji, M. Morifuji, K. Taniiguchi, C. Hamaguchi, T. Kunikiyo and M. Takenaka, J. Appl. Phys. **75**, 3500 (1994)

[23] Y. Wang and K. F. Brennan, J. Appl. Phys. **75**, 313 (1994)

[24] J. Kolník, Y. Wang, I. H. Oğuzman and K. F. Brennan, J. Appl. Phys. **76**, 3542 (1994)

[25] T. Kunikiyo, M. Takenaka, Y. Kamakura, M. Yamaji, H. Mizuno, M. Morifuji and C. Hamaguchi, J. Appl. Phys. **75**, 297 (1994)

[26] H. K. Jung, K. Taniguchi and C. Hamaguchi, J. Appl. Phys. **79**, 2473 (1996)

[27] I. H. Oğuzman, Y. Wang, J. Kolník and K. F. Brennan, J. Appl. Phys. **77**, 225 (1995)

[28] M. Reigrotzki, R. Redmer, I. Lee, S. Pennathur, M. Dür, J. F. Wager, P. Vogl, H. Eckstein and W. Schattke, J. Appl. Phys. **80**, 5054 (1996)

[29] I. H. Oğuzman, E. Bellotti, K. F. Brennan, J. Kolník, R. Wang and P. P. Ruden, J. Appl. Phys. **81**, 7827 (1997)

[30] S. M. Cho and H. H. Lee, J. Appl. Phys. **71**, 1298 (1992)

[31] V. M. Robbins, T. Wang, K. F. Brennan, K. Hess and G. E. Stillman, J. Appl. Phys. **58**, 4614 (1985)

[32] F. Osaka, T. Mikawa and T. Kaneda, IEEE J. Quantum Electron. **21**, 1326 (1985)

[33] F. Y. Juang, U. Das, Y. Nashimoto and P. K. Bhattacharya, Appl. Phys. Lett. **47**, 972 (1985)

[34] F. Osaka, T. Mikawa and O. Wada, IEEE J. Quantum Electron. **22**, 1986 (1986)

[35] T. Kagawa, Y. Kawamura, H. Asai, M. Naganuma and O. Mikami, Appl. Phys. Lett. **55**, 993 (1989)

[36] M. Tsuji, K. Makita, I. Watanabi and K. Taguchi, Appl. Phys. Lett. **65**, 3248 (1994)

[37] I. Watanabe, T. Torikai and K. Taguchi, IEEE J. Quantum Electron. **31**, 1826 (1995)

[38] V. M. Robbins, T. Wang, J. Tang, K. Hess, G. E. Stillman, R. J. McIntyre and P. Webb, IEEE Trans. Electron Devices **31**, 1977 (1984)

[39] C. A. A. ans S H Groves, Appl. Phys. Lett. **43**, 198 (1983)

[40] N. Tabatabaie, V. M. Robbins, K. F. Brennan, K. Hess and G. E. Stillman, IEEE Trans. Electron Devices **30**, 1608 (1983)

[41] T. P. Pearsall, R. E. Nahorny and J. R. Chelikowsky, Phys. Rev. Lett. **39**, 295 (1977)

[42] T. P. Pearsall, F. Capasso, R. E. Nahorny, M. A. Pollack and J. R. Chelikowsky, Solid State Electron. **21**, 297 (1978)

[43] F. Capasso, R. E. Nahorny and M. A. Pollack, Electron. Lett. **15**, 117 (1979)

[44] J. J. Berenz, J. Kinoshita, T. L. Hierl and C. A. Lee, Electron. Lett. **15**, 152 (1979)

[45] P. A. Wolff, Phys. Rev. **95**, 1415 (1954)

[46] W. Shockley, Solid State Electron. **2**, 35 (1961)

[47] G. A. Baraff, Phys. Rev. **128**, 2507 (1962)

[48] B. K. Ridley, J. Phys. C: Solid State Phys. **16**, 3373 (1983)

[49] M. G. Burt, J. Phys. C: Solid State Phys. **18**, L477 (1985)

[50] S. McKenzie and M. G. Burt, J. Phys. C: Solid State Phys. **19**, 1959 (1986)

[51] B. K. Ridley, Semicond. Sci. Technol. **2**, 116 (1987)

[52] J. S. Marsland, Solid State Electron. **30**, 125 (1987)

[53] L. V. Keldysh, Sov. Phys. JETP **10**, 509 (1960)

[54] M. V. Fischetti and S. E. Laux, Phys. Rev. B **38**, 9721 (1988)

[55] M. V. Fischetti, IEEE Trans. Electron. Devices **38**, 634 (1991)

[56] E. Cartier, M. V. Fischetti, E. A. Eklund and F. R. McFeely, Appl. Phys. Lett. **62**, 3339 (1993)

[57] J. Appl. Phys. **54**, 5139 (1983)

[58] E. O. Kane, Phys. Rev. **159**, 624 (1967)

[59] M. Stobbe, R. Redmer and W. Shattke, Phys. Rev. B **49**, 4494 (1994)

[60] C. J. Williams, PhD Thesis, University of Newcastle-upon-Tyne (1996)

[61] A. R. Beattie, J. Phys. C: Solid State Phys. **18**, 6501 (1985)

[62] A. R. Beattie, R. A. Abram and P. Scharoch, Semicond. Sci. Technol. **5**, 738 (1990)

[63] A. R. Beattie, R. A. Abram and P. Scharoch, Semicond. Sci. Technol. **7**, B512 (1992)

[64] S. P. Wilson, S. Brand, A. R. Beattie and R. A. Abram, Semicond. Sci. Technol. **8**, 1944 (1993)

[65] N. Sano, M. Tomizawa and A. Yoshii, Jpn. J. Appl. Phys. Part 1 **30**, 3662 (1991)

[66] T. Kunikiyo, M. Takenaka, M. Morifuji, K. Taniguchi and C. Hamaguchi, J. Appl. Phys. **79**, 7718 (1996)

[67] J. Allam, Jpn. J. Appl. Phys. Part 1 **36**, 1529 (1997)

[68] S. Imagana, K. Hane and Y. Hayafuji, J. Appl. Phys. **74**, 5859 (1993)

[69] P. D. Yoder, J. M. Higman, J. Bude and K. Hess, Semicond. Sci. Technol. **7**, B357 (1992)

[70] W. Quade, F. Rossi and C. Jacoboni, Semicond. Sci. Technol. **7**, B502 (1992)

[71] J. Bude, K. Hess and G. J. Iafrate, Semicond. Sci. Technol. **7**, B506 (1992)

[72] J. Bude, K. Hess and G. J. Iafrate, Phys. Rev. B **45**, 10958 (1992)

[73] M. J. Howes and D. V. Morgan, eds., *Gallium Arsenide: Materials, Devices and Circuits*, Wiley (1985)

[74] M. J. Adams and I. D. Henning, *Optical Fibres and Sources for Communications*, Plenum (1990)

[75] T. P. Pearsall, ed., *GaInAsP Alloy Semiconductors*, Wiley (1982)

[76] S. C. Jain, Advances in Electronics and Electron Physics, Supplement **24** (1994)

[77] D. T. Hughes, R. A. Abram and R. W. Kelsall, IEEE Trans. Electron Devices **42**, 201 (1995)

[78] K. Yeom, J. M. Hinkley and J. Singh, J. Appl. Phys. **80**, 6773 (1996)

[79] T. P. Pearsall, H. Temkin, J. C. Bean and S. Luryi, IEEE Electron Device Lett. **7**, 330 (1986)

[80] K. Yeom, J. M. Hinkley and J. Singh, Appl. Phys. Lett. **64**, 2985 (1994)

[81] J. R. Chelikowski and M. L. Cohen, Phys. Rev. B **14**, 556 (1976)

[82] S. Brand and R. A. Abram, J. Phys. C: Solid State Phys. **17**, L571 (1984)

[83] J. P. Walter and M. L. Cohen, Phys. Rev. B **5**, 3101 (1972)

[84] W. A. Harrison, *Solid State Theory*, McGraw-Hill (1970)

[85] F. Bassani and G. P. Parravicini, *Electronic States and Optical Transitions in Solids*, Pergamon (1975)

[86] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992)

[87] C. Jacoboni and L. Reggiani, Rev. Mod. Phys. **55**, 645 (1983)

[88] E. O. Kane, J. Phys. Chem. Sol. **1**, 249 (1957)

[89] M. L. Cohen and T. K. Bergstresser, Phys. Rev. **141**, 789 (1966)

[90] W. A. Harrison, *Pseudopotentials in the Theory of Metals*, Benjamin (1966)

[91] H. Ehrenreich, F. Seitz and D. Turnbull, eds., volume 24 of *Solid State Physics: Advances in Research and Applications*, Academic (1970)

[92] L. R. Saravia and D. Brust, Phys. Rev. **176**, 915 (1968)

[93] P. Löwdin, J. Chem. Phys. **19**, 1396 (1951)

[94] I. V. Aabrenkov and V. Heine, Philos. Mag. **12**, 529 (1965)

[95] A. O. E. Animalu and V. Heine, Philos. Mag. **12**, 1249 (1965)

[96] M. G. Burt, S. Brand, C. Smith and R. A. Abram, J. Phys. C: Solid State Phys. **17**, 6385 (1984)

[97] F. Stern, *Elementary Theory of Optical Properties of Solids*, volume 15 of *Solid State Physics: Advances in Research and Applications*, chapter 4, Academic (1963)

[98] O. Madelung, M. Schulz and H. Weiss, eds., *Landolt-Börnstein, Numerical Data and Functional Relationships in Science and Technology*, volume III/17a, Springer-Verlag (1982)

[99] J. Bude and K. R. Smith, Semicond. Sci. Technol. **9**, 840 (1994)

[100] P. T. Landsberg, *Recombination in Semiconductors*, Cambridge (1991)

[101] C. Kittel, *Introduction to Solid State Physics, 6^{th} edition*, Wiley (1986)

[102] A. R. Beattie and P. T. Landsberg, Proc. Roy. Soc. A **249**, 16 (1959)

[103] P. T. Landsberg, Proc. Phys. Soc. A **62**, 806 (1949)

[104] N. Sano and A. Yoshii, Phys. Rev. B **45**, 4147 (1992)

[105] M. Stobbe, A. Könies, R. Redmer, J. Henk and W. Schattke, Phys. Rev. B **44**, 11105 (1991)

[106] C. L. Anderson and C. R. Crowell, Phys. Rev. B **5**, 2267 (1972)

[107] A. R. Beattie, Semicond. Sci. Technol. **7**, 401 (1992)

[108] S. P. Wilson, S. Brand, A. R. Beattie and R. A. Abram, Semicond. Sci. Technol. **8**, 1546 (1993)

[109] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods*, Methuen (1964)

[110] D. T. Hughes, PhD Thesis, University of Durham (1989)

[111] N. Sano and A. Yoshii, J. Appl. Phys. **77**, 2020 (1995)

[112] N. Sano, T. Aoki and A. Yoshii, Appl. Phys. Lett. **55**, 1418 (1989)

[113] V. Chandramouli and C. M. Maziar, Solid State Electron. **36**, 285 (1993)